Stony Brook University

OFFICIAL COPY

The official electronic file of this thesis or dissertation is maintained by the University Libraries on behalf of The Graduate School at Stony Brook University.

© All Rights Reserved by Author.

Evolution of Complexity in Gene Regulatory Networks During Host-Parasite Coevolution

A Dissertation Presented

by

Jeewoen Shin

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Computational Biology)

Stony Brook University

December 2016

Stony Brook University

The Graduate School

Jeewoen Shin

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Thomas MacCarthy – Dissertation Advisor
Assistant Professor, Applied Mathematics and Statistics

Robert Rizzo - Chairperson of Defense
Professor, Applied Mathematics and Statistics

Sasha Levy
Assistant Professor, Applied Mathematics and Statistics

Joshua Rest
Associate Professor, Ecology and Evolution

This dissertation is accepted by the Graduate School

Charles Taber

Dean of the Graduate School

Abstract of the Dissertation

Evolution of Complexity in Gene Regulatory Networks During Host-Parasite Coevolution

by

Jeewoen Shin

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Computational Biology)

Stony Brook University

2016

Robustness, defined as tolerance to perturbations such as mutations and environmental fluctuations, is pervasive in biological systems. However, robustness often coexists with its counterpart, evolvability - the ability of perturbations to generate new phenotypes. Previous models of gene regulatory network evolution have shown that robustness evolves under stabilizing selection, but in more realistic scenarios such as coevolution, it may be advantageous to evolve sensitivity, i.e. for some mutations to change the phenotype. Furthermore, it is unclear how robustness and evolvability will emerge in common coevolutionary scenarios. In this dissertation, we consider three different two-species models of coevolution involving one host and one parasite population. First, we developed a two-population (host, parasite) model to investigate how robustness and evolvability become distributed within a network under antagonistic coevolution. We found that sensitivity follows a pattern, similar to that of the game "whack-a-mole", in which sensitive sites mutate, thus becoming insensitive, but new sensitive sites emerge to take their place. Second, we developed a host-virus interaction model focusing on

host resistance and viral pathogenicity which depend on quite different evolutionary conditions. Viruses may evolve cell entry strategies that use small receptor binding regions, represented by low complexity binding in our model. Our modeling results suggest that if the virus adopts a strategy based on binding to low complexity sites on the host receptor, the host will select a defense strategy at the protein (receptor) level, rather than at the level of the regulatory network - a virus-host strategy that appears to have been selected most often in nature. Lastly, we developed a model of the host innate immunity evolution in the context of host-virus coevolution. After viruses enter host cells, they interfere with innate immune systems via protein-protein interactions such as molecular mimicry of various host proteins involved in the immunity. We found that depending on different viral mechanisms for pathogenicity, hosts evolved to optimize the use of 1) mutations at protein-protein interaction sites to avoid mimicry and 2) environmental robustness in the innate immune systems imposed by viral disruption of the immune systems.

Table of Contents

List of Figures/Tables/Illustrations

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Abbreviations

GRN: Gene Regulatory Network

TF: Transcription Factor

PPI: Protein-Protein Interactions

SD: Standard Deviation

SEM: Standard Error of Mean

MVG: Modularly Varying Goal

ESBL: Extended-Spectrum Beta-Lactamase

SIS: Susceptible-Infected-Susceptible

S: The size of susceptible

I: The size of infected

TfR1: Transferrin Receptor-1

MMTV: Machupo virus

CAR: Coxsackie and Adenovirus Receptor

HCV: Hepatitis C virus

CCR5: CC-chemokine receptor-5

IFN: Interferon

HBV: Hepatitis B virus

IRF: Interferon regulatory transcription factor

IL: Interleukin

PKR: Protein kinase R

Acknowledgments

I would like to truly thank my advisor Professor Thomas MacCarthy, for his support, guidance, and advice during my Ph. D studies. Without his scientific guidance and warm encouragement, it would not have been possible to finish this valuable journey. I would also like to give special thanks my dissertation committee members Dr. Robert Rizzo, Dr. Sasha Levy and Dr. Joshua Rest for their time and careful consideration to detail. Friends and colleagues in MacCarthy lab have been always supportive and provided feedback on my research. I would like to thank my loving fiancé, Joo-won Kim who is also my closest colleague and academic mentor for encouraging me to pursue my dream. Lastly, but the most importantly, this journey would not have been even started without the support and encouragement of my parents, Mangyun Shin and Sukja Son. Thank you all for loving me, believing in me, financially supporting my study and inspiring me to pursue my dream. I also thank my brother, Wangsuk Shin, who cares and supports me like a big brother. I am also grateful to my grandparents, Soonja Park, Chang-gon Son, and Sowon Yeo for their love and heartfelt care.

Chapter 1. Introduction

1.1. Introduction

1.1.1. Robustness and evolvability in biological systems

Standing genotypic variation in biological systems occurs across many different scales ranging from coding region differences that affect amino acid sequences, through changes in metabolic pathways and signaling pathways, up to gene regulatory network changes affecting development and morphology. Organisms are also affected by environmental variation such as temperature changes, fluctuating concentrations of resources [1-4]. Although an environmental variation may affect standing genetic variation, mutations are, of course, the original source of variation [1, 5-7].

Robustness, defined as tolerance to perturbations such as mutations and environmental fluctuations, is pervasive in biological systems [5, 8]. Many studies have demonstrated the existence of robustness at many different biological scales including gene regulatory networks [9], RNA secondary structures [10], protein structures [11], signaling pathways [12-14] and metabolic networks[15]. Because of the long evolutionary time scales involved, experimental approaches to understanding the evolution of robustness in biological systems are extremely difficult. As an alternative, to address such questions, computational modeling and simulation approaches based on realistic representation of biological and chemical processes have been widely used for the last few decades. Early computational models of evolution aimed at understanding the relationship between gene-network evolution and behavior (gene expression dynamics) [16-18]. These studies found that, although a large number of different networks (genotypes) have the same gene expression dynamics (phenotype), they can usually be connected

to one another via minimal changes (e.g. creation or deletion of single *cis*-regulatory interactions) that might easily occur during evolution via mutation. This capacity for neutral evolution can facilitate the evolution of robustness since it allows a population to migrate towards more robust genotypes without altering the phenotype [19]. Numerous theoretical studies have shown that robustness will evolve in particular when the phenotype is under evolutionary pressure to remain constant (stabilizing selection). Experimental results are consistent with this notion. Gene networks in *E. coli*, for example, have been shown to be robust specifically to regulatory rewiring [20]. Similar experiments on metabolic networks, also in *E. coli*, have shown network robustness with respect to both gene knockouts and network rewiring [21-23].

However, robustness often coexists with its counterpart, evolvability - the ability of perturbations to generate new phenotypes. It has previously been suggested that evolvability can be facilitated by robustness. Mutations will tend to accumulate in populations with high robustness, leading to greater genetic variation, which in turn may facilitate access to new phenotypes [5, 19]. Work in the late 1990s on Drosophila Hsp90 (Heat Shock Protein 90, a chaperone targeting signal transducers) introduced the concept of phenotypic capacitance. A phenotypic capacitor is a mechanism which has the potential to expose the underlying genotypic variation. In the case of Hsp90 deletion, it was shown that an increased probability of stop codon read-through could expose cryptic variation in genotype that would not be translated in non-stress conditions, leading to increased phenotypic variation. Later work both theoretical and experimental showed more generally that environmental stress or stochastic processes could break robustness and drive phenotypic evolvability [24-26]. In this way, two opposite concepts, robustness and evolvability turn out to be tightly connected synergistic phenomena. While

genetic mutations are not often manifested at the phenotypic level in a robust system, they may accumulate silently and can drive phenotypic evolution by creating a wider range of mutational opportunities. Thus, perturbations in a phenotypic capacitor in which cryptic mutations are silently stored decrease robustness but facilitate evolvability and adaptation to environmental changes [5].

Many recent studies have shown that ecological interactions both within and between species, and particularly coevolutionary interactions, drive evolutionary changes on a far more rapid timescale than previously estimated [27-29]. Various forms of interaction occur among different species including mutualism, antagonistic coevolution and competition. In an ecological context, the interference within and among species is important to their evolution and survival [27-29]. Previous models of gene regulatory network evolution have shown that robustness evolves under stabilizing selection, but it is unclear how evolutionary features including robustness and evolvability will emerge in common but more complex coevolutionary scenarios. Although there have been many studies emphasizing importance of coevolution from both evolutionary and ecological perspectives [29], network model-based theoretical studies have not been used yet. Network modeling is a simple but powerful theoretical approach and has been widely used to improve our understanding of the evolution of diverse biological systems. It will be meaningful therefore, to use network models to understand how coevolutionary selection evolves networks and determines evolutionary properties such as robustness and evolvability [27]. Furthermore, as the two species interaction network that we consider here is eventually expanded to a bigger multiple species network, the network model might better explain the evolution of ecosystems [8].

1.1.2. Gene regulatory network evolution model

In traditional population genetics, a genotype-phenotype mapping is like a black box, and each genotype is assigned with a fitness value, 1-s, where s is selection coefficient. In contrast, computational models such as the gene regulatory network evolution model (also known as the Wagner's gene regulatory network model), combine a complex genotype-phenotype mapping (describing a gene regulatory network) with evolutionary dynamics. The Wagner model will be the basis for all the models presented here. An overview is shown in Fig 1.1.

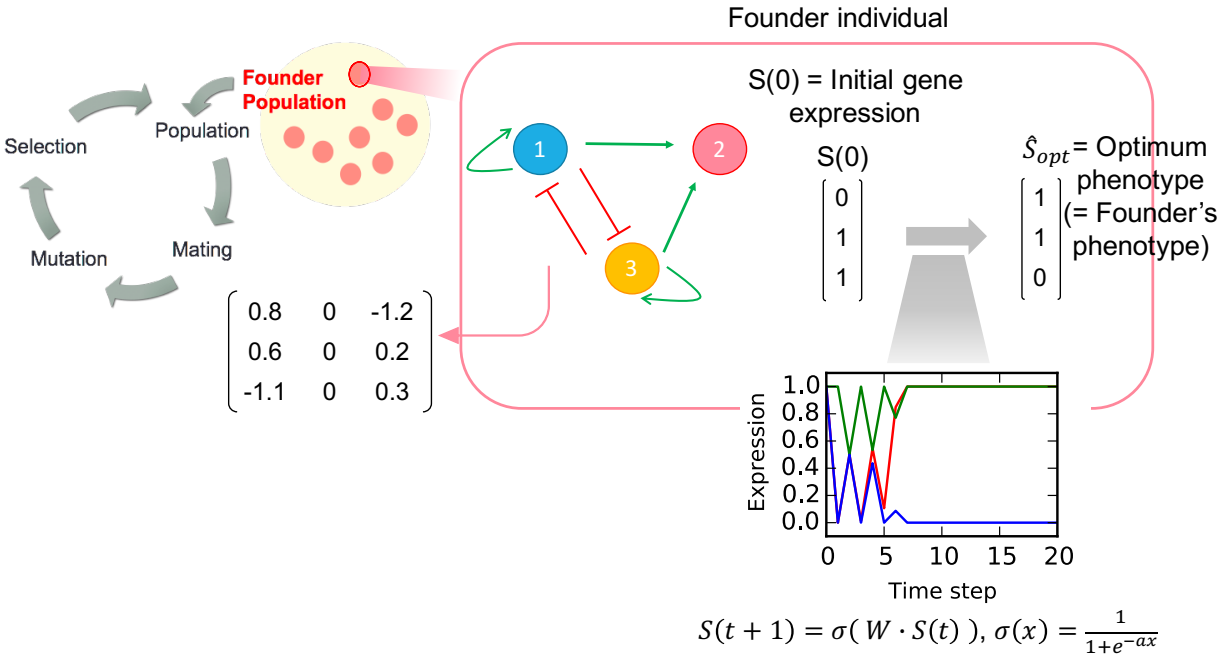


Figure 1. 1. Genotype-phenotype mapping and population level dynamics. A cycle of one generation includes a reproduction, a mutation and a selection step. In a homogeneous fixed-size founder population, a founder individual possesses a randomly assigned gene regulatory network with the initial gene expression. The stable gene expression (phenotype) can be obtained via gene expression dynamics. Under stabilizing selection, the founder's phenotype is used as the target (optimum) phenotype.

In the Wagner model, a genotype is represented as a matrix ($\boldsymbol{W}$) where $\boldsymbol{W(i,j) = w_{ij}}$ indicates i-th gene regulation by j-th gene product. Positive entries mean activation, negative entries mean repression and zeros indicate no interaction. Gene expression is represented as a vector ($\boldsymbol{S}$) of $\boldsymbol{N}$ genes. The initial gene expression is given as a random binary vector where 0 and 1 indicate unexpressed and expressed gene respectively. Gene expression at time $\boldsymbol{t+1}$ is $\boldsymbol{S(t+1) = \sigma\big(W \cdot S(t)\big)}$ where $\boldsymbol{\sigma(x) = \frac{1}{1+e^{-ax}}}$, and. $\boldsymbol{a(>0)}$ is a parameter which controls the steepness of the sigmoid function. When $\boldsymbol{S}$ reaches a steady state ($\boldsymbol{\widehat{S}}$: phenotype), the individual is used as a founder. The initial population is built using copies of the founder. In the reproduction step, each offspring is produced by inheriting each gene (i.e., a row in $\boldsymbol{W}$) from either parent. In the mutation step, the offspring's $\boldsymbol{W}$ is mutated with interaction additions, deletions and modifications. However, only offspring having high fitness value are likely to be selected to the next generation. The fitness function is $\boldsymbol{f\big(\widehat{S}\big) = e^{-\frac{D(\widehat{S},S^{OPT})}{N\alpha}}}$, where $\boldsymbol{S^{OPT}}$ is a phenotype of founder individual, $\boldsymbol{\alpha}$ is selection pressure and $\boldsymbol{D(S_1,S_2) = \sum_{i=1}^{N}\big(S_1(i) - S_2(i)\big)^2}$ ($\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are of the same length $\boldsymbol{N}$). Hence, offspring whose phenotype is similar to the founder's phenotype are most likely to be selected (stabilizing selection).

The gene regulatory network evolution model has been used previously to address a range of questions concerned with evolution of biological complexity [30, 31]. In previous studies, the original model has been extended to account for different system levels, including transcription factor (TF)-DNA binding interactions [32] and protein-protein interactions (PPI) [33] at the microscopic level, or, as presented below, between two different populations [34] at the macroscopic level. We were interested in considering how robustness and evolvability can evolve to be distributed across different system levels, depending on various model conditions.

In particular, we integrate protein-protein interactions (virus-receptor binding) and gene regulatory networks (which control receptor expression) in the context of an evolutionary model that represents both host and pathogen populations.

1.1.3. Research objectives

As described in section 1.1.2, the gene regulatory network evolution model has so far been mainly used to address the evolution of robustness under stabilizing selection in a single population. However, in more realistic scenarios different levels of systems interact each other and it is important to understand how interacting systems coevolve. Depending on interaction schemes, it may be advantageous to evolve sensitivity, i.e. for some mutations to change the phenotype instead of simply evolving mutational robustness. In this dissertation, we consider three different two-species models of coevolution involving one host and one parasite (virus) population.

In Chapter 2, we introduce our two-population (host, parasite) model and investigate how robustness and evolvability evolve within a gene regulatory network under antagonistic coevolutionary selection. In the model, parasites are modeled on species such as cuckoos where mimicry of the host phenotype confers high fitness to the parasite but lowers the fitness of the host. We study how sensitivity, defined as the potential to cause major phenotype changes, evolves and comes to be distributed in gene regulatory networks during antagonistic coevolution.

In Chapter 3, we introduce our model of host-virus coevolution involving two different levels of systems: a gene regulatory network and a protein-protein interaction. In this chapter, we focus on host resistance and viral pathogenicity which depend on quite different evolutionary conditions. We investigate model parameters that will encourage host individuals to evolve

network level resistance strategies to change receptor gene expression. We also explored conditions that the hosts will use to select a defense strategy at the protein (receptor) level, rather than at the level of the regulatory network - a virus-host strategy that appears to have been selected most often in nature.

In Chapter 4, we introduce our model of the host innate immunity evolution in the context of host-virus coevolution. After viruses enter host cells, they interfere with innate immune systems via protein-protein interactions such as molecular mimicry of various host proteins involved in the immunity. In this chapter, we investigate evolutionary features appear in host innate immune systems, and the conditions that induce evolution of gene regulatory network complexity. We discuss viral mechanisms for pathogenicity and how hosts evolve their defense mechanisms depending on different viral mechanisms.

Lastly, in Chapter 5, we summarize the three main studies in Chapter 2, 3, and 4, and propose future work.

Chapter 2. Antagonistic coevolution drives whack-a-mole sensitivity in gene regulatory networks

This chapter is adopted from the paper "Antagonistic Coevolution Drives Whack-a- Mole Sensitivity in Gene Regulatory Networks" [34].

2.1. Background

Robustness, defined as tolerance to perturbations such as mutations and environmental fluctuations, is pervasive in biology. Previous models of gene regulatory networks have shown that robustness can evolve when the phenotype is under evolutionary pressure to remain constant (stabilizing selection). But in more realistic scenarios such as coevolution, it may be advantageous to evolve sensitivity, i.e. for some mutations to change the phenotype.

Many recent studies have shown that ecological interactions both within and between species, and particularly coevolutionary interactions, drive evolutionary changes on a far more rapid timescale than previously estimated [27-29]. Here we use network modeling to understand how coevolutionary selection, rather than stabilizing selection, evolves network structure and function and how coevolution determines evolutionary properties such as robustness and evolvability [27]. We focus on a simple case of antagonistic coevolution between two populations, specifically a parasite population that uses mimicry of a complex phenotype as its survival strategy, as well as its host population. There are many documented cases of such interactions. A well-studied example is brood parasitism of cuckoos on their avian hosts. For instance, cuckoo finches (*Anomalospiza imberbis*) deposit their eggs in the nest of their host, the African tawny-flanked Prinia (*Prinia subflava*). By mimicking the eggshell morphology of their

hosts, the cuckoos trick their hosts into brooding these eggs. An evolutionary arms race between cuckoos and their host species drives continued variation in eggshell morphology in both species [35, 36]. In another example, coevolution of complex chemical signals occurs between *Maculinea alcon*, a parasitic butterfly species and their host, *Myrmica* ants [37]. *M. alcon* larvae emit a pattern of surface chemicals very similar to those of the ant larvae, leading the ants to adopt and feed the butterfly larvae as their own. An evolutionary arms race has arisen between these two species such that the ants evolve changes in their larval surface chemicals to discriminate their own larvae from those of the parasite whereas the parasite is continuously evolving to again produce a similar pattern.

It has previously been suggested that evolvability - the capacity for generating new phenotypes – can be facilitated by robustness, a somewhat counter-intuitive idea since evolvability and robustness would superficially appear to be opposite concepts [38]. However, with high robustness, mutations will tend to accumulate, increasing genetic variation in populations, which in turn may promote adaptation to new phenotypes [5, 19]. Phenotypic variation might be accessible during episodes of directional selection or particular conditions such as environmental stress [5, 6, 38-40]. Thus, under this model, periods of stabilizing selection allow genetic variation to accumulate, which is then eliminated by periodic selective sweeps and the cycle begins again with a new period of stabilizing selection. At the same time, the importance of this model remains unclear since few studies of network evolution have gone beyond stabilizing selection to investigate more realistic selection regimes [41]. Here we analyze host-parasite coevolution and find an entirely different strategy arises in which networks evolve a capacity for evolvability together with robustness against mutations. Here, evolvability in the network facilitates coevolutionary adaptation and is distributed throughout the network.

Previous studies have also shown there is a relationship between evolvability and modularity in networks. A strategy of using two target phenotypes presented, for example, in alternating succession has been used because it can select for distinct network modules, each of which is capable of generating one of the target phenotypes [42, 43]. One such study by Kashtan and Alon [42] used feed-forward logic networks and found that modularity evolved together with a fixed "evolvability node" which controlled the switch between two modules when mutated, thus switching phenotypes. Subsequent analyses showed evidence for modularity in other contexts including neural and metabolic networks [42, 44-46]. An alternative to a fixed "evolvability node" may be to have evolvability distributed throughout the network, allowing phenotype changes to occur in many different ways. Both types of evolvability are observed in nature [47]. Examples of fixed evolvability nodes include the Drosophila *shavenbaby* locus which predominantly controls trichome patterning [47], *Pitx1* which determines the pelvic spine phenotype in stickleback fish [48] and *optix* which controls rapidly evolving wing patterns in *Heliconius* butterflies [49]. Examples of distributed evolvability have been reported in bacterial and virus species including in *Helicobacter pylori* where a broad spectrum of genetic variations explains adaptation to its human host [50], in the pathogen *Pseudomonas aeroginosa* where antibiotic resistance evolves via several different mechanisms [51] and similarly in *E. coli* adaptation to low glucose environments [52]. Although the examples above illustrate the two extremes of what is likely a continuum between fixed and distributed evolvability, here we investigate a more general question - what conditions might favor the evolution of fixed vs distributed evolvability?

2.2. Model

2.2.1. Host-parasite coevolution model

The model largely follows previously published models [53-56] with the exception of selection, which here depends on interactions between the host and parasite populations. In our model, a genotype is represented as a matrix (W) where the elements w_{ij} describe the regulation of the $i-th$ gene by the $j-th$ gene product. Positive matrix entries represent activation, negative entries represent repression and zeros indicate no interaction. Gene expression is represented by a vector $S(t)$ containing elements $S_i(t)$ representing the expression level, in the range $(0,1)$, over time t of the $i-th$ gene. Initial gene expression, $S(0)$, is given as a random binary vector of 0 and 1 expression levels. Gene expression dynamics are determined by the difference equation $S(t+1)=\sigma(W\cdot S(t))$ where $\sigma(x)=\frac{1}{1+e^{-ax}}$ is a sigmoid function. The steady state gene expression, $\hat{S}$, is the phenotype and individuals not reaching steady state have zero fitness. The evolutionary simulation is initiated by creating a founder individual for each population in the form of a random matrix W of regulatory interactions containing non-zero elements with probability c, drawn from a Normal distribution, $N(0,1)$. The founder is copied to form the initial population of size M. Each population undergoes cycles of reproduction, mutation and selection. In the case of sexual reproduction candidate offspring are produced by inheriting a row in the matrix W at random from either parent. Here each row i represents regulatory interactions of the set of *cis*-regulatory elements controlling the expression of gene i. Row-wise inheritance implies inheritance of *cis*-regulatory regions and free recombination among loci. Under asexual reproduction, random parent genotypes are simply cloned. Following [54], mutations apply to the genotype of each offspring, W and may cause addition of new

network interactions (when element $w_{ij} = 0$) or deletions and modifications (when $w_{ij} \neq 0$). The mutation frequency per genotype is constant (μ). Mutations lead to either addition (ρ), deletion (ϕ) or modification (δ) of interactions. The addition and deletion rates are set to ensure that network density does not change from its initial value (See Parameters section below). In the selection step, the interaction between host and parasite populations determines a distinct fitness function for each population, as described in the following section 2.3.1.

2.2.2. Parameters

Unless otherwise stated, the simulation results used the following parameter values: number of genes, $N = 10$; population size, $M = 200$; mutation rate per genotype, $\mu = 0.1$; selection strength, $\alpha = 0.1$; asexual reproduction; network density, $c = 0.5$. As described previously [54], the network density c, will be at steady state when its difference in time, $\Delta c(t) = c(t) - c(t-1) = \mu\left(\alpha\left(1 - c(t)\right) - \phi c(t)\right) \big/ N^2$ is zero. We therefore chose the parameters for addition ($\rho = 0.025$) and deletion ($\phi = 0.025$) that satisfy $\Delta c(t) = 0$. Given these parameters, modifications are set to ($\delta = 1 - \phi$).

2.3. Results

2.3.1. Host-parasite coevolution model

To study gene regulatory network evolution under antagonistic coevolution we defined a model with two interacting populations. The model is an extension of a widely used single-

population model that assumes stabilizing selection [30, 55]. As in the previous model, each population functions at two broad levels: genotype-to-phenotype mapping and population dynamics (see Methods for details). For the genotype-phenotype mapping, the genotype is defined as a gene regulatory network of N genes represented by a $N \times N$ matrix, W, the entries w_{ij} of which represent the regulatory strength and sign of gene j on gene i ($N = 10$ was used for all results unless otherwise stated). The genotype is mapped to phenotype via gene expression dynamics. The gene expression levels at time t are represented by $S(t)$, a length N vector $S(t) = [s_1, s_2, \ldots, s_N]$ $(0 \leq s_i \leq 1, i = 1, \ldots, N)$. The genotype W defines a dynamical system that is used to determine steady state gene expression levels for each gene, which correspond to the phenotype, $\hat{S}$. Both host and parasite populations have a fixed number of individuals M. Cycles of reproduction, mutation and selection proceed in parallel as shown schematically in Figure 2.1a. Reproduction (either sexual or asexual) and mutation largely follow previous models [53, 54, 56]. Genotype mutations allow for creation and deletion of regulatory interactions as well as quantitative changes [54]. The main difference with previous models is at the selection stage, where the host and parasite populations interact by mutually determining fitness in the other population. To represent antagonistic coevolution in our model, we assume that a specific morphological pattern (e.g. egg surface color of cuckoos and hosts) is determined by a gene regulatory network, and the phenotype does not affect vitality and fertility. Hence, in the model, a candidate parasite individual has higher fitness when its phenotype is similar to that of a randomly chosen host individual (a new random host is chosen for each parasite at every selection step and similarly for each host). Thus parasite fitness is defined as: $f\left(\hat{S}_P\right) = e^{-\frac{D(\hat{S}_P, \hat{S}_H)}{\alpha}}$,

where $D(X,Y)=\frac{\sum_{i=1}^{N}(x_i,y_i)^2}{N}$, $\hat{S}_P$ is the parasite phenotype, and $\hat{S}_H$ is the phenotype of a randomly selected host individual. On the other hand, we assume the host has higher fitness when its phenotype is different from that of the parasite and therefore host fitness is defined as: $f(\hat{S}_H)=e^{-\frac{1-D(\hat{S}_H,\hat{S}_P)}{\alpha}}$ where $\hat{S}_P$ is the phenotype of a randomly selected parasite individual. α is a parameter representing selection pressure. The fitness functions are symmetric about $x=0.5$ (Figure 2.2) to avoid any bias in how selection is applied in host vs parasite. Although the initial phenotypes are random, this two-population approach allows the eventual target phenotypes to emerge from the model, in contrast to previous models where the target phenotypes are defined *a priori*. The fitness definitions used are analogous to the two examples of host-parasite evolutionary arms races described above (cuckoo finch and *M. alcon*) whereby similarity (and differences) in complex phenotypes are selected for: eggshell morphology in the case of the cuckoo finch or the pattern of larval surface chemicals in the case of *M. alcon*.

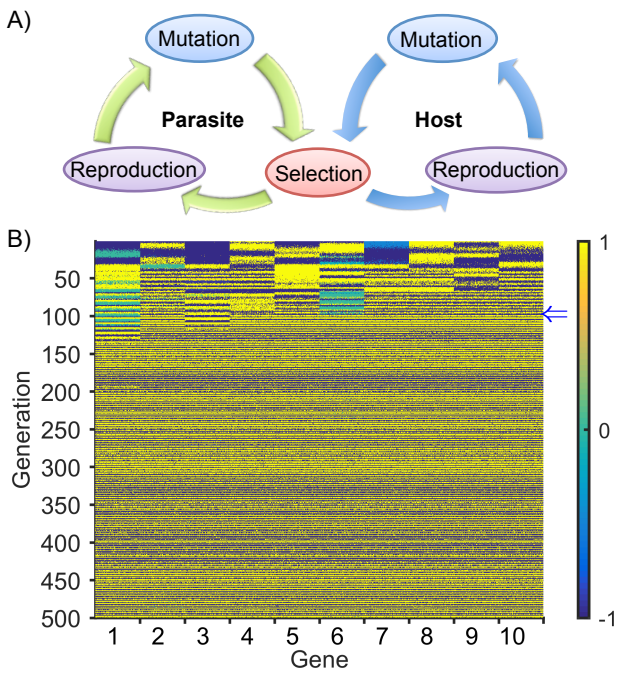


Figure 2. 1. Host-parasite model and alternating phenotype dynamics. A) Schematic overview of the host-parasite model. B) To compare host and parasite phenotypes, here in a typical simulation, gene expressions are rescaled from $[0,1]$ to $[-1,+1]$ so that for each gene the sign of their multiplied gene expression indicates whether their expressions are similar or different. Host and parasite phenotypes are compared, here in a typical simulation, by multiplying the expression of each gene, rescaled from $[0,1]$ to $[-1,+1]$, from one host and one parasite at each generation. In the horizontal direction, the leftmost block of columns represents the comparative expression (by multiplication of rescaled expressions) level of gene 1 for 200 host-parasite pairs (the pairings themselves are random). Similar gene expression between host and parasite is shown in yellow (parasite winning) and divergent expression in blue (host winning).

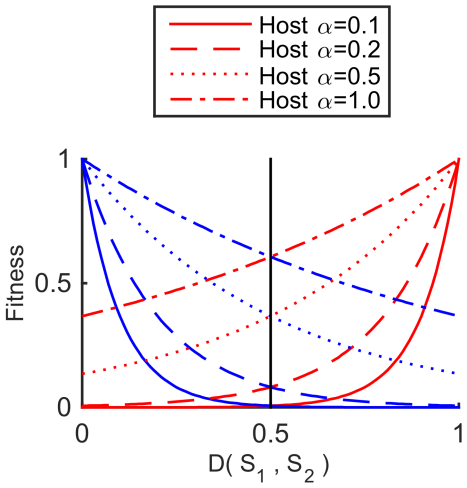


Figure 2. 2. Fitness function for host (red) and parasite (blue) for different selection pressure strengths. (α). $D(S_1,S_2)=\frac{\sum_{i=1}^{N}(S_{1i}-S_{2i})^2}{N}$ is the distance between two phenotypes S_1 and S_2. Host and parasite fitness values are symmetric about $D=0.5$.

2.3.2. Host and parasite populations evolve networks with distributed sensitivity and robustness

Under sufficiently strong selection pressure (α) both host and parasite populations reach a stage where their phenotypes alternate between one phenotype $\hat{S}$ and an approximately "inverted" version of the same phenotype, i.e. $1-\hat{S}=[1-s_1, 1-s_2, \ldots, 1-s_N]$. At a given generation, if the host population phenotype is $\hat{S}_H$ and that of the parasite is $\hat{S}_P=1-\hat{S}_H$, then the host will have high fitness and the parasite will have low fitness. However, if at a later generation the parasite population is able to "invert" its phenotype $\hat{S}_P \rightarrow 1-\hat{S}_P(=\hat{S}_H)$ and the host population maintains its phenotype ($\hat{S}_H$), then the parasite and the host phenotypes will

become the same - the host will now have low fitness and the parasite will have high fitness. The parasite population will continue "winning" until the host population is able to invert its phenotype, and the cycle continues. Figure 2.1b and Figure 2.3 Fig show this progression over time (vertical axis) for every gene expression level in every gene (horizontal axis) of every individual in a typical simulation (see Methods for parameter values used). Each cell in Figure 2.1b is colored blue when the expression level favors the host "winning" (i.e. when a host gene is on and the corresponding parasite gene is off and vice versa), and yellow if the parasite is "winning" (i.e. the host and parasite levels are the same). We see that by generation ~100 (blue arrow Figure 2.1b) both populations have converged to an alternating strategy as the rows alternate in color. Thus both host and parasite genotypes have become highly evolvable in response to phenotype changes in the opposite population. We are primarily interested in how these coevolutionary interactions between host and parasite populations affect gene regulatory network evolution and in particular how evolvability itself evolves within the networks. As expected, under weaker selection (approximately $\alpha > 0.15$, see Figure 2.5a) sensitivity did not evolve and the alternating phenotype was not observed. Hence, we focused here on the stronger selection case.

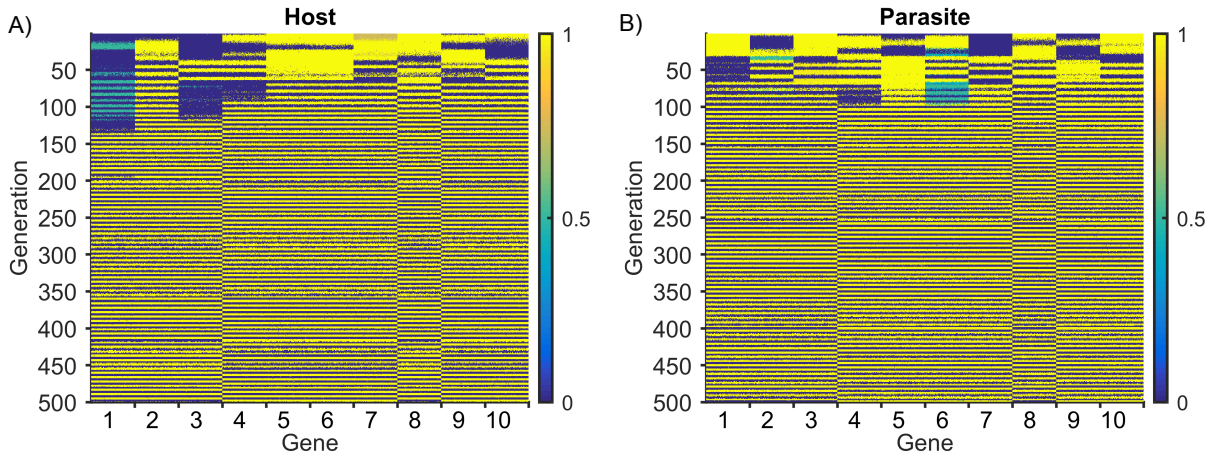


Figure 2. 3. Host and parasite populations used to generate Figure 2.1b. Here, as in Figure 2.1b, genes are shown on the horizontal axis and evolutionary time in generations on the vertical axis. The colors represent the gene expression levels of every gene in every individual, as indicated in the color bar.

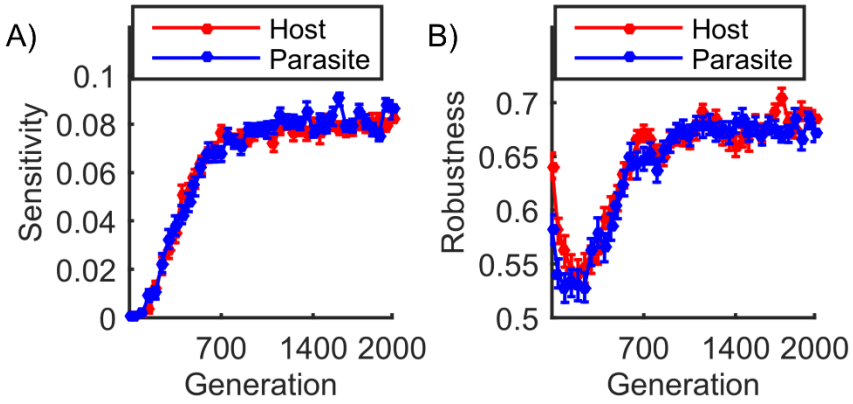


Figure 2. 4. Emergence of sensitivity. A) As coevolution proceeds, the sensitivity score (***SS***) increases monotonically reaching a plateau in both host and parasite. B) Robustness in the remaining (non-sensitive) part of the network was defined as the fraction of mutations that leave the phenotype unchanged if we exclude phenotype inversions (see main text). Both plots show mean values for 100 simulations with the error bars indicating standard error of the mean (SEM).

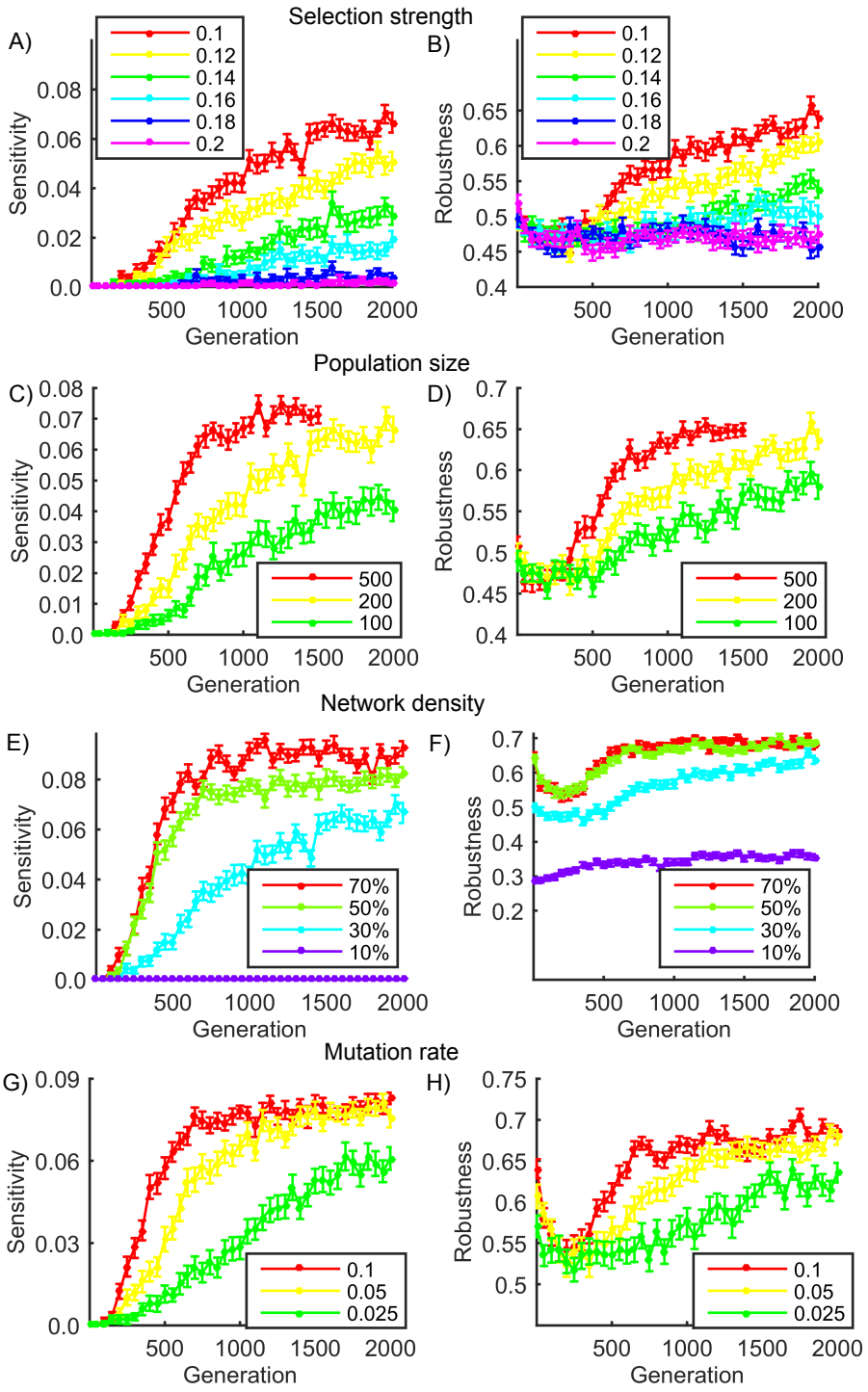


Figure 2. 5. Effect of parameter changes on the evolution of sensitivity and robustness. Results analogous to Figure 2.4 (for sensitivity and robustness) for varying parameters of the model. In each case, we only vary one parameter, maintaining the others fixed. (A) and (B) are for different values of the selection strength, α ($c = 0.3, M = 200, \mu = 0.1$); (C) and (D) are for

population size, M ($c = 0.3, \sigma = 0.1, \mu = 0.1$); (E) and (F) for network density, c ($\sigma = 0.1, M = 200, \mu = 0.1$); (G) and (H) are for mutation rate, μ ($c = 0.5, \sigma = 0.1, M = 200$).

We next sought to identify the mechanism underlying the phenotype inversion process, i.e. the evolution of evolvability. One possibility is that the alternating phenotype strategy would evolve in the form of a particular "evolvability hotspot" or interactions in the network, analogous to those identified previously by Kashtan *et al*. [42] in modular networks. A mutation in an "evolvability hotspot" would be highly likely to cause a phenotype inversion. An alternative scenario is one in which the capacity for phenotype inversion is highly distributed, and phenotype inversion can occur in many different places throughout the network, albeit with low probability. To assess these effects we implemented two measurements: first, a sensitivity score (SS) that estimates the overall probability that a mutation will cause a phenotype inversion (see Methods), and secondly a measure of how distributed the sensitivity is within the set of network interactions that can cause a phenotype inversion, as described below. In addition, to measure the effects of coevolution on the remaining parts of the network (that do not cause phenotype inversions) we also quantify mutational robustness in this subset of network interactions.

Figure 2.4a shows the progression of the sensitivity score (SS) during a typical simulation. Here we see that at the beginning of the coevolutionary process, because host and parasite networks are random, both have a negligible number of sensitive interactions and the mean SS is close to zero. As antagonistic coevolution proceeds and both populations evolve towards the alternating phenotype strategy, they both acquire sensitive interactions and the mean SS increases, eventually reaching a plateau. For the set of parameters shown in this example (see Methods), SS reaches approximately 0.08. Although this qualitative behavior is observed across a large range of parameter values, there are quantitative differences. Thus, the steady state SS

level is reduced, as expected, if selection pressure is lower (Figure 2.5a) and with smaller population sizes (Figure 2.5c) where random drift effects are greater. Also, networks with a greater density of connections can evolve sensitivity more easily (Figure 2.5e). Lastly, note that multiple simultaneous mutations can occur within a single genotype, particularly when the frequency of the single mutation is high, as is often the case when a population is undergoing a phenotype inversion (Figure 2.6a, b). Although such events occur at low frequency, we found that, in cases of double mutations at least one of the mutation positions had a high sensitivity score whereas the other usually had a sensitivity score that was either very low or zero (Figure 2.6c). Thus, a phenotype inversion is most often achieved with a single point mutation at a sensitive interaction, although occasional double mutations where at least one mutation is at a sensitive interaction can also cause a phenotype inversion.

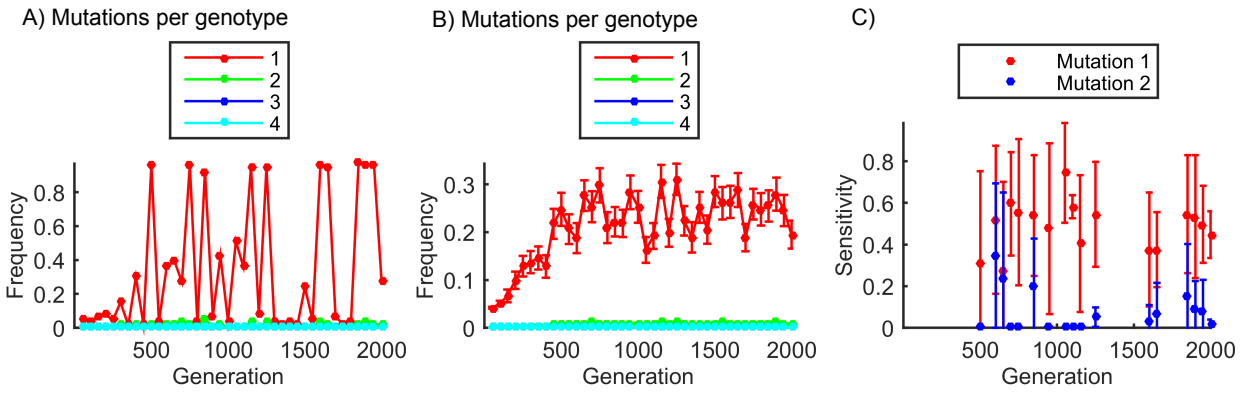


Figure 2. 6. Analysis of cases with multiple mutations. In the case of (A) a single typical simulation, and (B) averaged over 100 simulations, we compared the genotype of each individual with the ancestor genotype by back-tracking asexually reproducing populations. Curves show the frequency of single (red) and multiple (2 (green), 3 (blue), and 4 (cyan)) mutations over time. Error bars represent one SEM. (C) For the same simulation shown in (A), we measured the sensitivity score at those interactions that mutated when there were two mutations. The higher of the two sensitivity scores is shown in red, and the lower of the two is shown in blue. The error bars represent one SD.

The sensitivity score (SS) estimates the probability of causing a phenotype inversion. We

found that the capacity for causing a phenotype inversion is distributed across a large number of sensitive network interactions, and we therefore sought to quantify how sensitivity was distributed throughout the network. Sensitivity might either be distributed fairly equally among these interactions, or unequally in the sense that particular interactions are likely to cause a phenotype inversion whereas others interactions do so only with low probability. To quantify the distribution of sensitivity we first chose the subset of network interactions, w_{ij}, that exhibit sensitivity, where the subset is defined as those having a (interaction specific) sensitivity score $SS_{ij} > 0$ (see Methods). We compared the observed standard deviation (SD) of the SS_{ij} values to the SD of a null model that assumes the observed total sensitivity in this set of nodes is randomly distributed (see Methods). We consistently found that the null model has a comparable, and even slightly higher, variance of sensitivity within the sensitive interactions than the evolved networks (Figure 2.7a). Thus the levels of sensitivity are at least as similar amongst themselves than would be expected by chance given the observed total sensitivity in the network (see Methods).

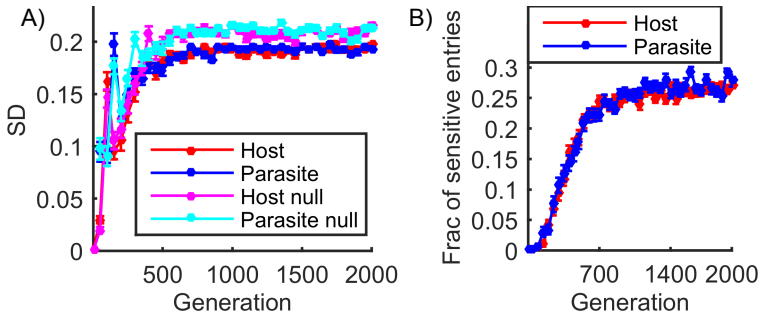


Figure 2. 7. Emergence of sensitive interactions and distribution of sensitivity score (*SS*) among the sensitive interactions. (A) Standard deviation (SD) of sensitivity scores at sensitive interactions for which $SS_{ij} > 0$, in red for host, blue for parasite. Null model results (see Methods) are also shown for host in magenta and for the parasite in cyan. The observed SD is comparable and even slightly below the SD of the null model. (B) As coevolution proceeds, the fraction of sensitive interactions in the network for which $SS_{ij} > 0$ increases monotonically reaching a plateau in both host (red curve) and parasite (blue curve).

Apart from causing a phenotype inversion, a mutation may either (a) leave the phenotype unchanged, which indicates robustness, or (b) cause the phenotype to change only partially which will usually be sub-optimal. As a measure of robustness, Figure 2.4b shows the fraction of mutations that leave the phenotype unchanged if we exclude phenotype inversions, i.e. $(a)/\big((a)+(b)\big)$. These results show that robustness initially decreases but then increases, eventually reaching a level higher than that of the initial population.

Note that the initial host and parasite populations have random phenotypes, generally their phenotypes are not in either similar or inverted forms. In addition, sensitivity does not exist before coevolution. Therefore, during the initial phase, all host/parasite individuals are under evolutionary pressure to explore alternative phenotypes to counter the other (parasite/host) population, which is also in a similar situation. Partial phenotype changes will therefore be beneficial until both populations enter the process of phenotype inversion. This is why robustness decreases in the earliest stages of coevolution (Fig 2-4b). However, once the capacity for phenotype inversion has evolved, partial phenotype changes will not be beneficial especially under strong selection and there is selection pressure for mutations to either preserve or invert the phenotype. This is why robustness increases together with sensitivity, and why robustness eventually exceeds the initial (pre-selection) levels. Again, we found that the phenomenon of increased robustness is observable across a wide range of parameter values although the range is more limited under sexual reproduction than it is with asexual reproduction (Figure 2.8). Generally though, robustness evolves in the parts of the network that are not causing phenotype inversion. Thus in the steady state, both robustness and evolvability coexist in the network under coevolutionary selection.

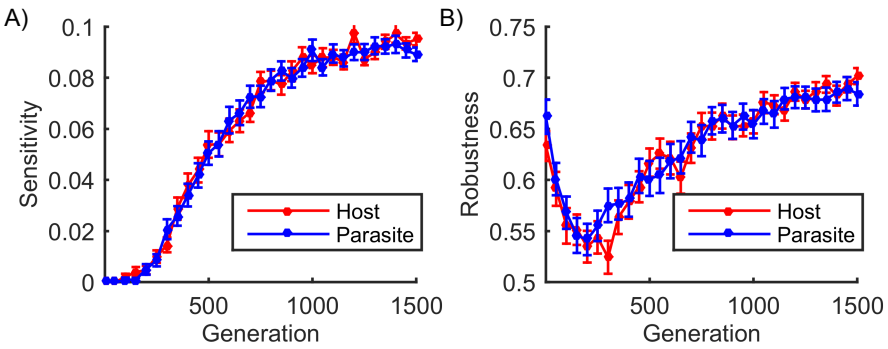


Figure 2. 8. Evolution of sensitivity and robustness under sexual reproduction. The results for sexual reproduction are qualitatively equivalent to those for asexual reproduction. However, for many parameter combinations in which robustness evolves under asexual reproduction, it does not evolve to be higher than the initial (random) case under sexual reproduction. Here, in plot (B) we show an example (parameters: $c = 0.7$, $M = 500$, $\alpha = 0.1$) for which the robustness clearly evolves.

Although we observe that mutational robustness evolves under antagonistic coevolution, environmental robustness appears to coevolve to a much lesser extent. Previous studies have shown that even without direct selection for environmental robustness, mutational and environmental robustness will coevolve under stabilizing selection [57, 58]. Environmental robustness was evaluated via perturbations of the initial gene expression levels and then by measuring the phenotypic distance between the perturbed and unperturbed cases (see Methods). Given the trend for mutational robustness (Figure 2.4b), the overall pattern was similar to that expected (Figure 2.9). However, the phenotypic distance increased to steady state levels that were well above those observed initially, indicating an overall reduction in environmental robustness. This was the case regardless of whether the perturbation rates were low or high relative to the mutation rate.

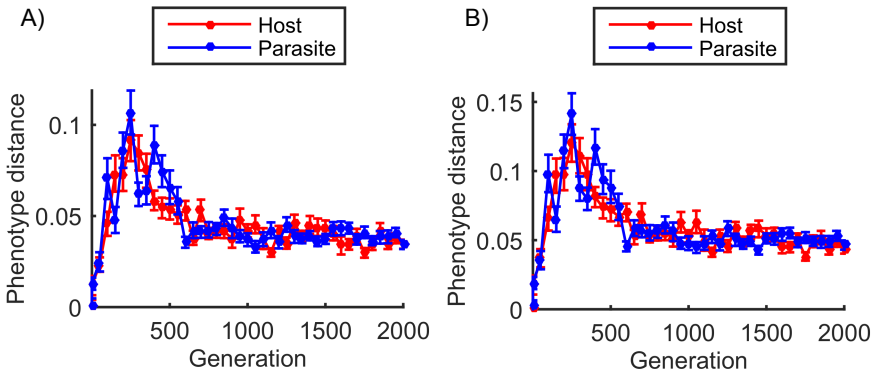


Figure 2. 9. Progression over time of phenotype distance in response to perturbations of the initial conditions to evaluate environmental robustness. The initial gene expression levels were perturbed 500 times for each individual in the population (A) at a rate 0.01/gene and (B) 0.2/gene. The phenotype distance was used (see Methods) to evaluate the environmental robustness.

We have addressed the simple case of equal population sizes for host and parasite. This case is relevant to many real host-parasite interactions such as the example of the cuckoos and their avian hosts discussed above, where the populations appear to be relatively stable and of comparable size [58]. Clearly however, host and parasite populations will often differ in size. We therefore evaluated the case of host population size = 100 and parasite population size=1000 (and vice-versa), finding only slight differences with the case of equal population sizes (Figure 2.10 vs. Figure 2.4). However, due to computational constraints we were unable to model much larger population sizes and we therefore leave a more thorough evaluation of unequal population sizes for future work.

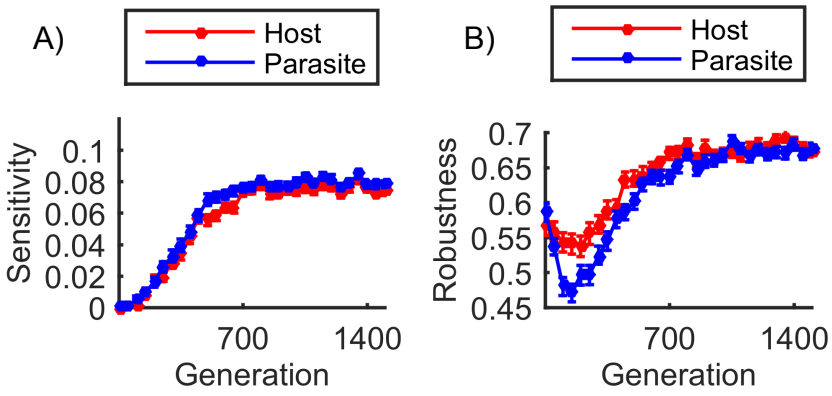


Figure 2. 10. Sensitivity and robustness over time for asymmetric population sizes. (A) Sensitivity and (B) robustness over time for asymmetric population sizes. We tested one order of magnitude difference between host population size = 100 and parasite population size = 1000. These plots are in the same format as for Figure 2.4; all other parameter values are the same as for Figure 2.4.

2.3.3. Sensitive regulatory interactions are highly labile throughout evolution

We next investigated whether sensitivity is preserved at particular points in the network or whether it changes over time. As described above for the case of modular networks, sensitivity will often evolve to be focused on "hotspots" that control distinct phenotypes and which do not change over time [42, 44]. To assess the changes in the sensitive interactions over time we used asexual reproduction. Under asexual reproduction, tracing the ancestral lineage is straightforward because there is a single parent for each individual, and after G generations each individual in the population needs G ancestral genotypes to store its genetic history. In contrast, under sexual reproduction each individual needs at most $2^1+2^2+\cdots+2^G$ ancestral genotypes, which rapidly becomes unwieldy. We consider the set of sensitive points of the network (i.e. those interactions w_{ij} with sensitivity score $SS_{ij}>0$ that may cause a phenotype inversion) and how this set changes over time. We selected networks at a particular steady state

generation and compared these to ancestral networks at various evolutionary distances. The comparison was done by measuring the similarity, in terms of sensitivity, between the ancestral and derived networks using the Jaccard index (see Methods) as shown in Figure 2.11a. Given that the phenotype is constantly changing, to ensure a valid comparison we only compared with ancestral networks having the same phenotype. As shown in Figure 2.11a, the overlap in sensitivity remains high only for a short time period, before dropping almost to levels that would be expected by chance (null model Figure 2.11a – also see Methods). However, at steady state the sensitivity remains stable, as do the total number of sensitive interactions (generations ~1000 onwards, Figure 2.4a and Figure 2.7b). Thus, sensitive interactions are highly labile and on average, each time a sensitive interaction is eliminated by mutation, a new one emerges to take its place. Colloquially this property is known as "whack-a-mole", named after the fun park game, and we therefore refer to this phenomenon as whack-a-mole sensitivity.

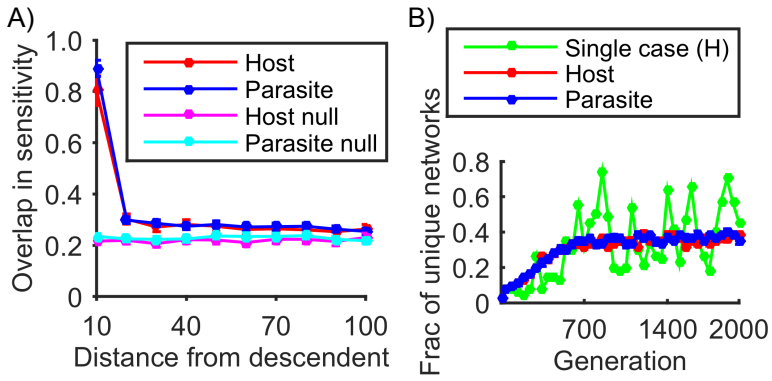


Figure 2. 11. Lability of sensitive interactions and network diversity. A) Comparison of the sensitivity network interactions from a single individual in a population at generation 2000 with its ancestors using the Jaccard index to quantify the overlap in sensitivity (see Methods). (B) Time course of network diversity, defined as the number of distinct networks, simplified to sign (-1/0/+1) form, and expressed as a fraction of the population (green line: host population of single simulation). Apart from the green line in (B), both plots show mean values for 100 simulations with the error bars indicating SEM.

Even though sensitive interactions are labile and are constantly being relocated, we thought there might be a specific subset of interactions with consistently high sensitivity. Alternatively, there might be no persistence in the sensitive interactions or any such interactions would be rapidly lost. Consistent with the latter scenario we found there are no interactions with a significantly high frequency of being a persistent sensitive interaction within a population and throughout a simulation, as shown in Figure 2.12. Figure 2.12a shows, for a typical simulation, the frequency at which each interaction w_{ij} was sensitive over a period of 1500 generations while sensitivity and robustness were at steady state levels. Figure 2.12b shows the change in sensitivity over time for two particular interactions in Figure 2.12a (those that had the highest and lowest overall sensitivity respectively). Figure 2.12c shows the same data in histogram form (green curve) together with the mean value for many simulations (red curve). Even though there appears to be no preference for particular positions within the matrix, we tested whether there was a higher-level preference for particular rows of the interaction matrix W, which represent the *cis*-regulatory elements for each gene. For this, we considered the total sensitivity score for each row (i), SS_i, and in particular, tracked the row i_{max} for which the value of SS_i is maximal within each individual (Figure 2.13a). We found that rarely does a particular i_{max} dominate both the population and throughout generations (Figure 2.13b). We repeated this analysis for columns, which represent gene outputs regulating genes, finding similar results. Thus, there does not appear to be any predilection for sensitivity to be associated with particular genes.

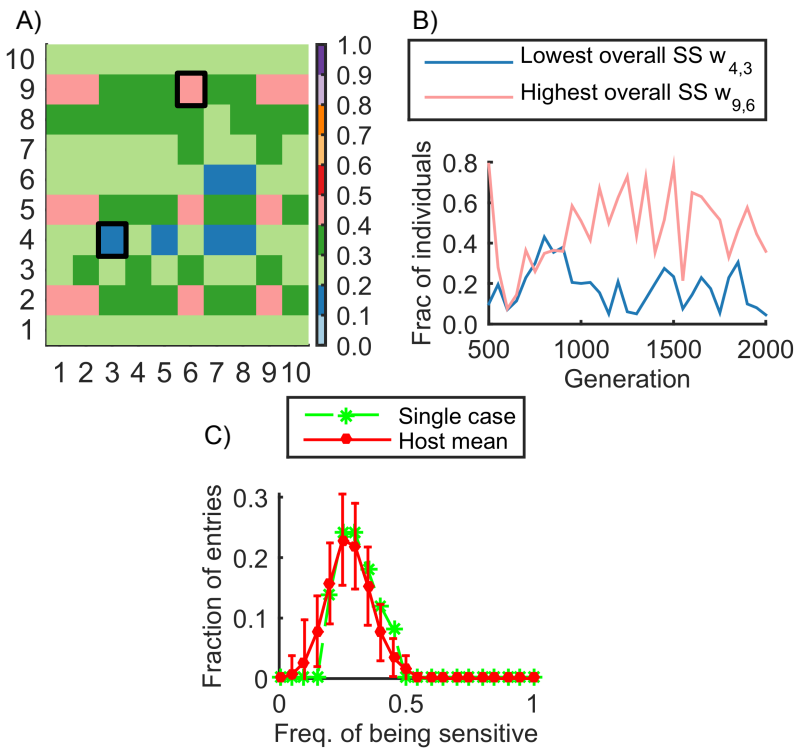


Figure 2. 12. Distribution of sensitivity throughout the network. A) Frequency of being a sensitive interaction in the $N \times N$ matrix of interactions (with $N = 10$) in a typical simulation. From generation 500 onwards we identified the sensitive gene interactions w_{ij} ($SS_{ij} > 0$), then measured the frequency for each w_{ij} being sensitive within the population, at intervals of 50 generations. We sum the frequencies over time and normalize to the interval $[0,1]$ as indicated by the colors. Generally, there are no interactions that appear to dominate within each population over many generations. B) Detailed progression of sensitivity over time for two particular interactions in (A). These interactions had the lowest (blue) and highest (pink) overall sensitivity, as indicated by the black squares in (A). C) Distribution of the frequency of being sensitive in all $N \times N$ interactions for all host individuals (green dashed line: the host population of (A), red solid line: mean of 100 simulations). Since distributions are mostly right-skewed there are no interactions that dominate in terms of sensitivity. Error bars indicate one SD.

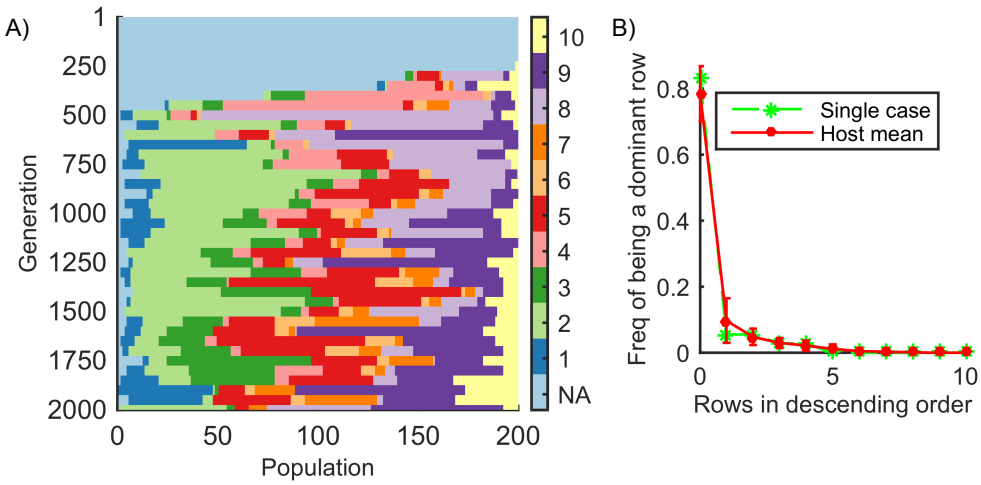


Figure 2. 13. Distribution of higher-level (row) sensitivity in the population. A) Every 50 generations we determined the sum of sensitivity scores (SS_{ij}) for each row in every individual in the host population of a typical simulation. In particular, we tracked the row i_{max} for which the value of SS_i is maximal within each individual. The row i_{max} of each individual is indicated by a different color, and the profile of for the entire population is represented by a row. We sampled these values every 50 generations (vertical axis). Light blue was used (NA on color bar) if there are no sensitive interactions in that particular network. For example, in generation 2000, shown at the bottom of the plot, 53 out of 200 individuals had $i_{max} = 2$. We define a "dominant" row as existing when more than half of the population has the same i_{max}. Thus for example, at generation 500, row $i_{max} = 8$ is dominant because more than half the individuals in the population have $i_{max} = 8$ (light purple). Row 10 (beige) on the other hand, is never dominant. The green curve in plot (B) shows, in rank order, the frequencies with which i_{max} was dominant for each generation in plot (A). In most cases there was no dominant row and we classified these cases as "row 0". For this analysis we considered only populations in steady state, i.e. from generation 500 onwards. For example, in plot (A), row 4 and 8 were dominant in 5.56% of the generations for each, more than any other row (rank #1), and this is shown in plot (B) as green dots at (1,0.0556) and (2,0.0556). The red curve shows the mean values for 100 independent simulations.

2.3.4. Antagonistic coevolution drives high levels of diversity

As explained in the introduction, another way by which phenotypic innovation has been

proposed to occur is through increased genetic variation, which is promoted by robustness. However, if a sensitivity mechanism has evolved to generate the appropriate phenotype changes, it does not, in principle, require high levels of genetic variation to function. To investigate the observed levels of genetic variation in the population that has evolved sensitivity we used a measure that simplifies each network using the sign of each matrix entry $\mathrm{sgn}(w_{ij})$, then counts the number of distinct (simplified) networks, expressed as a fraction of the population. Figure 2.11b shows how this diversity measure increases over time. Taking a typical host case (green curve) as an example we found that in the final population there were 91 distinct networks, which expressed as a fraction of the total population, leads to a diversity measure of 91/200=0.45. The average trend (red and blue curves) shows diversity increasing over time, eventually reaching a plateau. This diversity is a consequence of the beneficial mutations (occurring at sensitive interactions) being broadly distributed throughout the network, thus making multiple evolutionary pathways available. Taking this analysis further, we used the same diversity metric to measure the level of variation generated by stabilizing selection (see generations 1-500 in Figure 2.14) and found that the level of diversity was consistently below that observed under antagonistic coevolution. Although the comparison needs to be interpreted cautiously given that stabilizing and coevolutionary selection are quite different, we include it here to emphasize the high degree of diversity observed under coevolution. This high diversity occurs despite there being, in principle, no requirement for it.

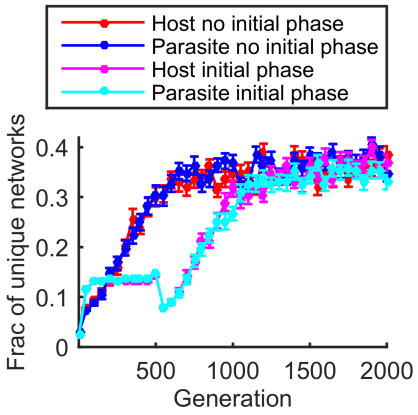


Figure 2. 14. Evolution of diversity with an initial phase of stabilizing selection. As described in the main text, we measure diversity in the population as the fraction of distinct networks in the population (simplified to sign form: +1/0/-1). We performed simulations with an initial phase of 500 generations under stabilizing selection (magenta and cyan curves) and include here the original results without the initial phase for comparative purposes.

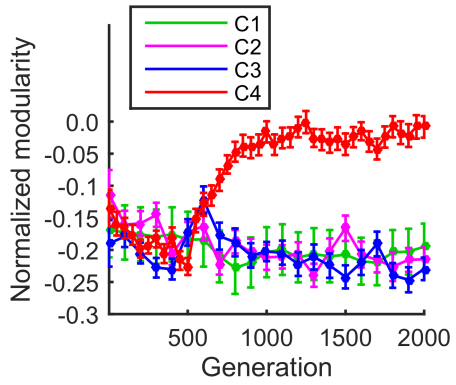
Lastly, we assessed the impact of having an initial phase of stabilizing selection that allows each population to evolve robustness and accumulate genetic variation independently before the coevolutionary process begins. As shown in Figure 2.14 (generation 500 onwards), genetic variation increases under the initial stabilizing selection phase (previous model of [55] was used for this), reaching a plateau by generation ~100. Once coevolution begins, at generation 500, genetic diversity is reduced as both populations pass through a bottleneck but then increases eventually exceeding the level achieved under stabilizing selection. However, the dynamics are not significantly different from those observed without the initial phase of stabilizing selection, and therefore the initial phase appears not to offer any advantage.

2.3.5. Innovation arising from sensitivity does not require modularity

As mentioned in the Introduction, previous studies [42, 44, 46, 57] have investigated the

conditions leading to increased modularity in a network using multiple target phenotypes. In each case, the target phenotypes contained a combination of features such that a modular network evolves. In particular, the study by Espinosa-Soto and Wagner [43] used two target phenotypes that only overlapped partially, leading to increased modularity (Figure 2.15a, red curve). Applying the same modularity measure to our own simulations (see Methods), we found that increased modularity did not evolve for any combination of parameters. We thought this might be because in our model, the entire phenotype alternates, in contrast to only part of the phenotype in the Espinosa-Soto model. However, a variant of our model in which only half the phenotype genes participate in host-parasite fitness and the other half of them are under stabilizing selection also did not evolve modularity (Figure 2.15a, green curve). The other key differences between the two models are how gene interactions and expression levels are represented (real vs discrete), the method of presentation of target phenotypes (sequentially alternating vs simultaneous) and the type of perturbation (mutational vs environmental), as summarized in Figure 2.15b. We therefore tested variant models that contained mixtures of features from either of the models but were unable to find increased modularity for any of the variant models (Figure 2.15a). These results suggest that modularity will evolve only under the very specific conditions. Biologically, the most important of these conditions is perhaps the nature of the perturbations, which can broadly be interpreted as growth-related or developmental for our model vs physiological or environmental in the case of the Espinosa-Soto model.

A)



B)

	Host-Parasite GRN coevolution model (M1)	Modularity evolving GRN model (M2)	C1	C2	C3	C4
Populations	Host and parasite	One population	M1	M1	M1	M2
A gene interaction	Represented as a value following normal distribution	Either 0 (no interaction), -1 (repression) or 1 (activation)	M1	M2	M2	M2
Gene expression	Real value between 0 and 1	Either -1 (unexpressed) or 1 (expressed)	M1	M2	M2	M2
Target	Alternating single target determined by host and parasite populations which is not given to a simulation	Fixed multiple simultaneous targets given to a simulation	M1	M1*	M1*	M2
Stage	Developmental process	Physiological process	M1	M1	M2	M2
Environmental perturbation	Environmental perturbations are not considered on initial gene expression	Environmental perturbations on initial gene expression	M1	M1	M2	M2

Figure 2. 15. Evolution of modularity in different model variants. A) The table describes features (1st column) that are different between our model (M1, described in 2nd column) and the Espinosa-Soto model (M2, described in 3rd column). Using the measure defined in Espinosa-Soto and Wagner, we measured modularity in our model (curve C1 in plot) and reproduced the results of model M2 (curve C4). We further tested two variant models that had features of both models, as indicated in columns 5 and 6, which correspond to curves C2 and C3 in the plot. The variant models did not show increased modularity over time. In these simulations, coevolution begins at generation 500 for all models C1 ~ C4. Initial network density, $c = 0.3$ for all four models to match the parameters used in Espinosa-Soto and Wagner. To make the models comparable, for models C2 and C3, we adopted the convention in the M1 model of defining only half the genes using the opposite population (either host or parasite) as a reference phenotype for defining fitness. The remaining genes used the founder individual, as in the stabilizing selection model without antagonistic coevolution, as in initial phase of Figure 2.14.

Another relevant study, by Kashtan and Alon [42] also found that modularity evolved in

the network together with persistent sensitive nodes. This was achieved using alternating target outputs, so-called Modularly Varying Goals (MVGs), which were defined as pairs of logical functions containing different combinations of sub-goals. For example, the authors defined two functions (of 4 inputs, X, Y, Z and W) as G1: (X XOR Y) OR (Z XOR W), G2: (X XOR Y) AND (Z XOR W). Although the model used was based on logical circuits and therefore quite different to the one we have used here, we evaluated whether using this particular pair (G1, G2) of MVGs would also result in long-term sensitive nodes. We implemented this using a single population model with networks having 4 designated input genes and 6 interacting regulatory genes, one of which is considered the output. Since there are $2^4 = 16$ possible inputs, fitness was defined as the fraction of correct input-output mappings. We evolved the population using alternating targets (G1, G2, G1, …) for 50 generations per target. For a population evolved under one target (e.g., G1), we assessed sensitivity, and in particular we considered any mutated network as sensitive if it matched the alternate target (e.g., G2) in more than 12 out of 16 input-output pairs (i.e., a fraction of 0.75). The threshold was set to 0.75 because we did not observe the average fitness exceeding this level for either target (Figure 2.16a). As shown in the Figure 2.16b, we do observe that a subset of persistent sensitive network nodes evolves; this is the subset of nodes with frequency of sensitivity equal to 1. However, in contrast to the previous study we did not observe increased modularity over time (Figure 2.16c), presumably because most sensitive interactions are not persistent, but highly transient, with frequencies of sensitivity between zero and one (Figure 2.16b). To further confirm that it is indeed the MVGs that facilitate the appearance of the subset of persistent sensitive nodes, we checked two further scenarios using the single population model. Firstly, we used two alternating targets in which half of the target genes ($N/2$) are kept the same as the founder phenotype and the other half are

inverted every 50 generations. This model is similar to that of Espinosa-Soto described above, except that the targets alternate in time, rather than being selected for simultaneously. In the second case, we simply alternated between the founder phenotype and its inverted form, again every 50 generations. In neither of these cases did we observe the emergence of persistent sensitive nodes (Figure 2.17) as we observed with the MVGs.

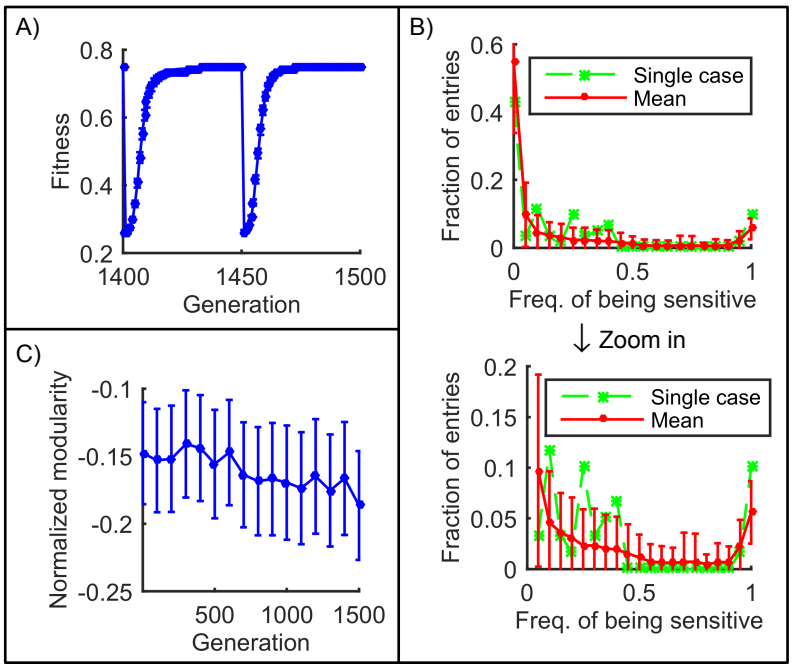


Figure 2. 16. Persistent/dominant sensitive interactions appear under selection for repeatedly switching Modularly Varying Goals (MVGs). (A) Fitness vs time for MVGs that switch every 50 generations. Fitness drops when the goal is changed and reaches equilibrium within approximately 20 generations. (B) Distribution of the frequency of being a sensitive interaction among all $N \times N$ interactions. This figure is the equivalent of Figure 2.12c for the case of MVGs. The bottom figure presents the same data, but has been zoomed in by omitting the left-most data point (fraction=0). The non-zero tail, and especially those interactions that have frequency of being sensitive =1, shows there are persistent sensitive interactions. (C) While persistent sensitive interactions do appear under the MVG model as shown in (B), modularity does not evolve because labile sensitive interactions are still present in these networks.

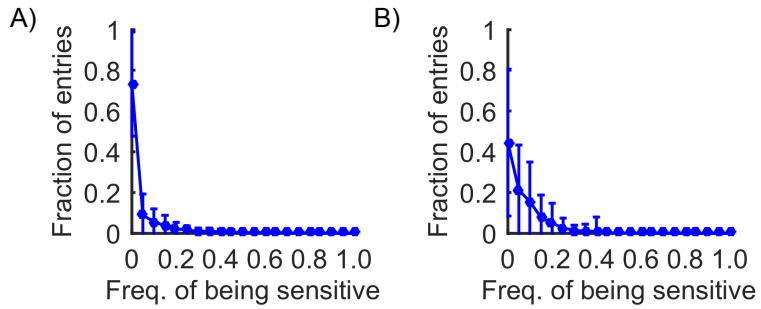


Figure 2. 17. Distribution of the frequency of interactions being sensitive among all $N \times N$ interactions for a single population under alternating selection strategies. The format is equivalent to Fig 4C. As explained in the main text, in (A) we used two alternating targets in which half of the target genes (N/2) are kept the same as the founder phenotype and the other half are inverted. In (B) we simply alternated between the founder phenotype and its inverted form. In both cases, switching between the two target goals occurs every 50 generations.

A second important difference between our approach and that of Kashtan and Alon lies in the mutation model. The Kashtan and Alon study used only topology changes, whereas our approach allows for both quantitative interaction modifications and topology changes. To investigate this difference further, we used our model to evaluate differences in the contribution of weight modifications vs topology changes. We found that increasing the relative importance of topology changes (by increasing the parameters for addition, ρ, and deletion, ϕ) did not qualitatively change our results and in particular, did not create persistent sensitive nodes in the network (Figure 2.18a, b). A reduction in the relative use of topology changes (by reducing ρ and ϕ) also did not change results qualitatively (Figure 2.18c, d). In conclusion, these analyses suggest that MVGs explain the major difference in outcomes, namely persistent evolvability nodes in the Kashtan and Alon model compared to distributed and labile evolvability nodes in our host-parasite coevolution model.

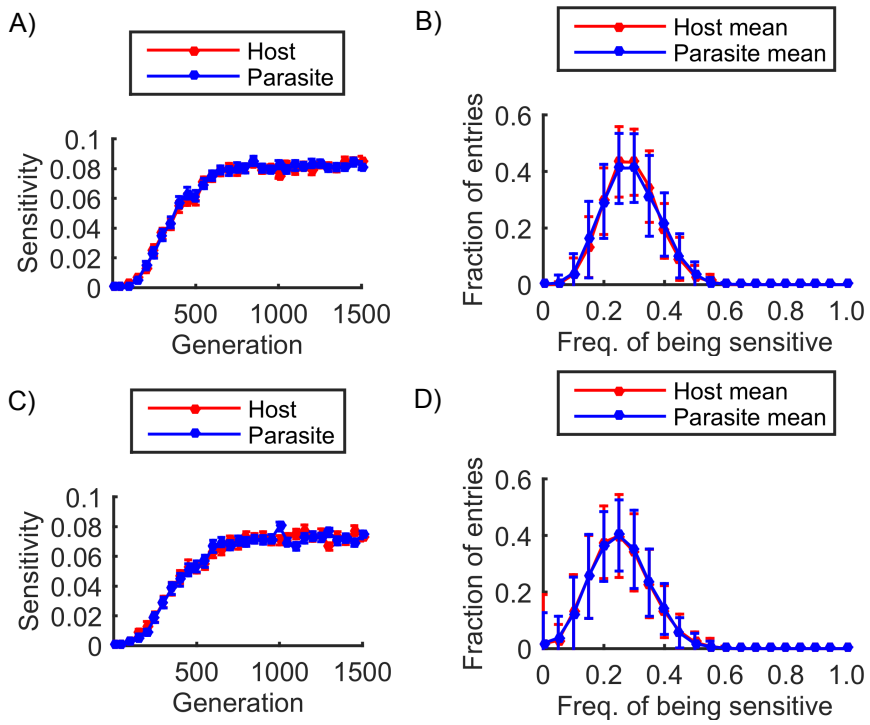


Figure 2. 18. Emergence of sensitivity (A, C) and the distribution of the frequency of interactions being sensitive among all $N \times N$ interactions (B, D) for different addition (ρ) and deletion (ϕ) rates. The format is equivalent to Figures Figure 2.4a and Figure 2.12c respectively (where $\rho = \phi = 0.025$). (A) and (B) are results for a 2.5X higher addition/deletion rate of $\rho = \phi = 0.0625$, whereas (C) and (D) are for the 2.5X lower addition/deletion rate of $\rho = \phi = 0.01$. Other parameters remain as described in Methods (section "Parameters").

2.4. Methods

2.4.1. Sensitivity score

As described above, a mutation is defined as the replacement of one element w_{ij} ($i, j = 1,\ldots,N$) with a random number drawn from a Gaussian distribution, $N(0,1)$ if the interaction is either modified or added and with zero if the interaction is deleted. The sensitivity score is calculated by estimating the expectation of a phenotype inversion given a random

mutation. This involves evaluating whether a mutation that would change $w_{ij} \to l$ will generate a phenotype inversion ($k(l)=1$) or not ($k(l)=0$). Because the probability of the mutation $w_{ij} \to l$ follows a continuous Gaussian distribution $f(l)$, we employ a discrete approximation given by evaluating $f(l)$ at $2L/\delta+1$ positions across the range $[-L\ ,\ L]$ separated by small intervals of size δ. More formally, the sensitivity score of an interaction (w_{ij}) in a network is measured as $SS_{ij}=\sum_{l=-L}^{L}\delta\cdot f(l)\cdot k(l)$ where $l\in\{-L+n\cdot\delta \mid -\mathrm{L}+n\cdot\delta\le L,\ n\in\mathbb{Z}^*\}$ (= the range of mutation: $w_{ij}\to l$), $f(l)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{l^2}{2\sigma^2}}$ (normal distribution probability density function with mean=0) and $k(l)=1$ if the phenotype is inverted by the perturbation $w_{ij}\to l$, otherwise $k(l)=0$. We consider a phenotype as inverted if the L_1 distance ($\|X-Y\|_1=\sum_{i=1}^{N}|x_i-y_i|$) between the original phenotype and a perturbed phenotype by the $w_{ij}\to l$ mutation, excepting those N_b genes that have basal expression $(s_i=0.5)$ due to not having inputs, is greater than $p_{flip}\cdot(N-N_b)$. For all results reported we used $p_{flip}=0.9$ and $\delta=0.02$, $\sigma=1$ and $L=3$, which covers the range of 99.73% of possible mutations at w_{ij}. The SS of a genotype (W) is the average of SS_{ij} for all elements w_{ij}, i.e., $SS=\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{SS_{ij}}{N^2}$.

2.4.2. Lability of sensitive interactions

The overlap in sensitivity, which describes the similarity of two genotypes, u and v, is measured as the Jaccard index, $J(u,v) = \frac{|A_u \cap A_v|}{|A_u \cup A_v|}$, where A_u is the set of sensitive interactions in u for which $SS_{ij} > 0$ and A_v is the set for genotype v. We calculate $J(u,v)$ for an individual (u) and its ancestor (v) of the same phenotype. Comparing with an ancestor of the same phenotype is fairer than comparing with an inverted phenotype. Because it is not possible to guarantee that an ancestor at a particular previous generation will have the same phenotype, we chose the closest ancestor having the same phenotype within a window of size 10 (i.e., assuming intervals of size 100, ancestors in ranges of 1-10,11-20, .., 91-100 generations previous). As a null model, sensitive interactions, with $SS_{ij} > 0$ are randomly redistributed in the networks. The overlap in sensitivity for the null model is calculated in the same way and the mean overlap for 100 null models is used (Figure 2.11a).

2.4.3. Null model for distribution of sensitivity in sensitive interactions

Assuming that a network has x interactions with sensitivity score $SS_{ij} > 0$ and the sum of these x sensitivity scores is H. For the null model we used a standard string cutting method that generates x numbers such that their sum equals H. This was implemented by choosing $x-1$ random numbers in the range $(0, H)$ and then calculating the distances between the adjacent numbers including the end points 0 and H. These distances (of which there are x), are random numbers which are distributed according to a Dirichlet distribution, and whose sum is H. As we did for the original network, we calculated the standard deviation (SD) of these x randomly

distributed sensitivity scores. We repeated this process 100 times and compared the mean value with the SD of the original network (Figure 2.7a).

2.4.4. Measuring environmental robustness

To quantify environmental robustness, we perturbed initial gene expression 500 times for each individual in the population by changing $s_i \to 1 - s_i$ at a rate 0.01/gene (Figure 2.9a) and 0.2/gene (Figure 2.9b). We then calculated the phenotype distance $D\left(S_1, S_2\right) = \frac{\sum_{i=1}^{N} \left| S_{1i} - S_{2i} \right|}{N}$ between unperturbed (S_1) and perturbed (S_2) phenotypes excluding phenotype inversion cases as the measure of environmental robustness.

2.4.5. Evolution of modularity under coevolutionary selection

The modularity measure we used is taken from [59]. We restate the definition as follows: Given a graph $G\left(V, E\right)$, where $\left|V\right| = k, \left|E\right| = l,$ the vertices of $G\left(V, E\right)$ can be clustered into n clusters, $C = \left\{C_1, C_2, \ldots, C_n\right\}, 1 \le n \le k$. Modularity is defined as

$$Q\left(C\right) = \sum_{i=1}^{k} \left[\frac{\left|E\left(C_i\right)\right|}{l} - \left(\frac{\sum_{v \in C_i} \deg\left(v\right)}{2l} \right)^2 \right]$$ where $E\left(C_i\right) = \left\{ \left\{v_h, v_t\right\} \in E \mid v_h, v_t \in C_i \right\}$ and

$\deg\left(v\right) = \left| \left\{ \forall v_t \ne v \mid \left\{v, v_t\right\} \in E \right\} \right|$.

We adopted the (Espinosa-Soto) model described in [43] to our antagonistic coevolution

model as follows. The previous study used two target phenotypes simultaneously, each of which has a conserved part and a distinct part as input gene expressions. To emulate this in our model, we assigned half the genes to be under stabilizing selection and the other half under coevolutionary selection. The set of genes under stabilizing selection has the same target phenotype throughout the simulation whereas the other is under coevolutionary selection. To represent environmental perturbations, the input gene expression is perturbed by changing $s_i \to 1-s_i$ at a rate $0.15/N$ as described in [43], for 400 experiments. The fitness function of an individual is $f = 1-e^{-3\gamma}$, $\gamma = \sum_{i=1}^{400}\left(1-D_i/D_{max}\right)^5/400$, where D_i is the Hamming distance between the target phenotype and a new phenotype from $i\ th$ perturbed input. For the model version that assigns half of the genes to be under stabilizing selection and the other half to be under coevolutionary selection but does not include environmental perturbations, we calculate two types of fitness for stabilizing and coevolutionary selection respectively:

$f_s = 1-e^{-\frac{d}{\alpha}}$, $d = \sum_{i=1}^{N_s}\left(S_{optimal}(i)-\hat{S}(i)\right)^2 \Big/ \left(N_s\cdot\zeta\right)$ and

$f_c = \begin{cases} e^{-\frac{1-d}{\alpha}}, & host \\ e^{-\frac{d}{\alpha}}, & parasite \end{cases}$, $d = \sum_{i=1}^{N_c}\left(S_{antagonist}(i)-\hat{S}(i)\right)^2 \Big/ \left(N_c\cdot\zeta\right)$.

$\zeta = 1$ for continuous expression levels and $\zeta = 4$ for discrete (-1,+1) expression levels. N_s is the number of genes under stabilizing selection and N_c is the number of genes under coevolutionary selection. Survival requires both f_s and f_c to exceed a uniformly-distributed random value in the range $[0,1]$.

2.5. Discussion and conclusion

Previous models of gene regulatory networks have shown that mutational robustness evolves under conditions of stabilizing selection [55, 57]. However, under more realistic scenarios, such as coevolution, evolvability may be advantageous. It is unclear though, how sensitivity and robustness will evolve and in particular how they will become distributed throughout a regulatory network. To investigate this, we developed a two-population (host-parasite) model of antagonistic coevolution. Although previous studies [42-44] had investigated the evolution of sensitivity in networks under fluctuating environmental conditions, a key novelty of our model is that the fitness landscapes are emergent properties of the inter-population interactions. This approach avoids the need to impose a changing environmental regime externally. Furthermore, the pace of evolution is dictated largely by the model's ability to adapt. Self-contained models such as these represent a step towards open-ended evolutionary models that will be critical in the longer term to understanding how biological complexity evolves.

We found that sensitivity increases after the initiation of coevolution and becomes highly distributed throughout the network. At the same time, the remaining (non-sensitive) parts of the network evolve to become robust. Interestingly, genetic diversity evolves to be higher under antagonistic coevolution than under stabilizing selection. There are two obvious sources of diversity in this case. Firstly, in the non-sensitive parts of the network, robustness facilitates the accumulation of genetic variation via a well-understood mechanism [60]. Secondly, because sensitivity is distributed across the network, there are many different ways in which mutations cause phenotype inversions, contributing to diversity particularly after several rounds of selection. If sensitivity were not distributed, but were focused on a particular "evolvability

hotspot", genetic diversity could in principle be far lower.

We found that robustness evolves in the parts of the network that are not involved in phenotype inversion. Interestingly, this robustness evolves more easily under asexual reproduction (Figure 2.4b) than sexual reproduction (Figure 2.8b). Generally speaking, we found that under sexual reproduction, a combination of higher network density, stronger selection pressure and/or larger population size was required in order to attain levels of robustness comparable to the asexual case, suggesting that recombination load is having nontrivial effects under sexual reproduction. In support of this, theoretical population genetic studies investigating the evolution of recombination [61, 62] have shown that asexual reproduction will be favored over sexual reproduction under antagonistic coevolution when the two modes are allowed to compete. At the same time, a previous study using a similar network model to ours [63], but having a single population under conditions of stabilizing selection, demonstrated that recombination load evolves to minimal levels under stabilizing selection. Thus it would appear that recombination load evolves to be higher under antagonistic coevolution than under stabilizing selection.

We found that coevolutionary selection drives networks to evolve labile sensitivity such that evolvability and robustness are continuously redistributed throughout the network. Sensitive points within the network cause a phenotype inversion when mutated (from $\hat{S} \to 1 - \hat{S}$), but the mutation by definition also changes the genotype, in particular by causing a change $w_{ij} \to w_{ij}'$. Assuming the lineage continues through another phenotype inversion (from $1 - \hat{S} \to \hat{S}$) and w_{ij}' is not mutated again, then w_{ij}' will most likely no longer be sensitive. However, each time a sensitive point in the network is "used up", a new sensitive point emerges elsewhere, thus

maintaining overall sensitivity at approximately constant levels. We refer to this process as whack-a-mole sensitivity, named after a fun park game in which targets are removed from one place only to reappear elsewhere. A comparable whack-a-mole process also appears to occur with meiotic recombination hotspots in mammals [64, 65]. During meiosis, recombination breakpoints are frequently initiated at DNA motif hotspots recognized by the PRDM9 protein. However, DNA repair mechanisms cause hotspots to be preferentially lost in the gametes of heterozygote (hotspot/non-hotspot) individuals and the net effect is for recombination hotspots to be lost over time. However, by means that are still not well understood, the overall number of hotspots (in humans for example) remains approximately constant while the positions of recombination hotspots are transient and vary within humans [66, 67], suggesting there must be a mechanism for generating new hotspots to replace those that have been lost, i.e., a whack-a-mole process. Broad distributions of mutations have been observed in antibiotic resistance, for example in bacteria which produce extended-spectrum beta-lactamase (ESBL) enzymes [68, 69]. In this case, many distinct point mutations occurring in ESBL genes such as TEM-1 and SHV-1 transform the active site of the enzyme. More than 330 ESBL variants including TEM- and SHV- type variants have been reported [69]. Whack-a-mole sensitivity may explain these rapidly expanding mutations in genes encoding ESBLs, thus helping to predict the evolution of resistance.

In our model the ongoing phenotypic inversions are dependent on successive mutations that accumulate across many different loci. Because we found that sensitive nodes are labile (whack-a-mole sensitivity), this means that over time similar mutations at a particular gene regulatory interaction might have distinct phenotypic effects. For example, a phenotype inversion ($S \to 1-S$) might be caused by a mutation at a particular sensitive site w_{ij}. The

original phenotype S might then be restored by a reverse mutation at the same site. However, because sensitivity is distributed across the network, the restored phenotype S is more likely to arise through a mutation at some other site different than w_{ij}. Indeed, several generations may pass before this reverse mutation occurs and by that time other mutations may have accumulated in the network. In this new genetic background, the "reverse" mutation may no longer have the same effect. This is a clear example of serial epistasis - the dependency of mutational effect on the genetic background established by previous mutations [70]. A widely-cited study of serial epistasis in a natural population involves the evolution of resistance to the insecticide diazinon in populations of Australian sheep blowfly [71, 72]. Here, an early resistance mutation arose conferring higher fitness in the presence of insecticide, but lower fitness compared to wildtype in the absence of insecticide. A second mutation then evolved to ameliorate the deleterious mutation, thus restoring fitness to wildtype levels for the double mutants.

A key issue in evolutionary biology is understanding the extent to which epistasis, and in particular serial epistasis, determines the path of evolutionary change [70]. Such evolutionary constraints have been shown clearly at the level of individual proteins, for example, in a classic study of the evolution of novel function in vertebrate steroid receptors [73], the authors evaluated experimentally the inferred ancestral proteins leading to the separate evolution of mineralocorticoid and glucocorticoid steroid receptors. They found that structural interactions imposed constraints that determined a specific ordering for the observed evolutionary substitutions. At the same time, the importance of serial epistasis in larger-scale systems such as regulatory networks is less well understood [74]. Our results suggest that whack-a-mole sensitivity will evolve as an emergent property of the network when there is distributed sensitivity and the serial epistasis effects that come with it.

Taken together, under conditions of strong antagonistic coevolution, sensitivity in gene regulatory networks evolves to be broadly distributed and highly labile. Our results suggest there will be no central network elements that determine phenotype changes in the long term. Previous studies had found that network modularity could evolve in the context of alternating environments comparable to those that emerge in our model [42, 43]. A modular network architecture can facilitate phenotype switching by perturbing key interaction(s) between modules [46]. However, we observe an entirely different mechanism based on sensitivity in which modularity does not play a role. When we compared with the Espinosa-Soto model [43], we found that modularity did not evolve even when we adopted many model features, and perhaps the most relevant difference with that model lies in the nature of the perturbations (Figure 2.15). Modularity may be more likely to evolve in the face of environmental perturbations than in networks faced predominantly with mutational perturbations [45, 46, 75]. When we compared with the Kashtan model [42], we found that introducing Modularly Varying Goals (MVGs) could, to some extent, drive the evolution of persistent sensitive interactions (Figure 2.16b), although network modularity did not increase (Figure 2.16c). As we have observed, distributed sensitivity offers the advantage of allowing a large number of mutations throughout the network to generate phenotype changes. If a network has many different regulatory interactions that enable rapid adaptation via point mutations, the network does not have to mutate a specific interaction back-and-forth in order to repeat the process.

Chapter 3. Potential for evolution of complex defense strategies in a multi-scale model of virus-host coevolution

This chapter is adopted from the paper "Potential for evolution of complex defense strategies in a multi-scale model of virus-host coevolution" [76].

3.1. Background

Viruses and their hosts engage in evolutionary arms races in the form of continuous molecular level changes that determine the mechanisms of infection and defense [77-80]. The evolutionary dynamics are determined in large part by host susceptibility and viral pathogenicity and ultimately depend on molecular interactions between genes and their products [81-83]. These relentless evolutionary arms races drive genetic diversity in both host and pathogen [78, 84, 85]. More generally, host-pathogen interactions have been proposed as a major factor in the evolution of biological complexity [34, 86-88].

If we consider humans and other higher organisms as potential hosts, they will usually evolve at much slower rates than the viruses that infect them [89]. At the same time these hosts are highly complex organism and will usually have far greater resources in terms of potential defense mechanisms and, more generally, in terms of genetic information to deal with the viral infections. Viral entry will commonly involve binding interactions with receptors on the host cell surface [90, 91]. Most host cells will have a large number of cell surface receptors, many of which are involved in essential functions such as detection of signaling molecules (e.g. hormones) or nutrients, but which can be usurped by viruses as cell entrance mechanisms [92, 93]. Functional redundancy among receptors is common. For example, nectins are cell entry

receptors of Herpes simplex virus (HSV) and are involved in cell adhesion. Functional redundancy within the nectin family and also other cellular adhesion proteins can compensate for particular nectins [94]. Also, in humans there are 19 known chemokine receptors which activate the same chemokine signaling pathway but some of these have highly specific receptor binding ligands whereas others may bind multiple ligands [95]. Interestingly, some viruses produce mimics of chemokine receptor binding ligands, or may encode their own chemokines and chemokine receptors [96]. For example, CCR5 and CXCR4 act as co-receptors for HIV-1 entry [97], and the Respiratory Syncytial Virus (RSV) produces its own version of the chemokine CXC3 which binds to the host receptor CX3CRI, thus facilitating RSV infection [98].

While there are multiple mechanisms of infection and resistance across many levels, virus entry into the host cell is the first and essential step that must succeed for a viral infection to proceed [90, 91]. Thus, preventing virus entry has often been the preferred strategy for therapeutic development [90, 99, 100]. On evolutionary timescales, hosts can evade receptor-mediated viral entry in several ways including amino acid changes at the binding sites to inhibit protein interactions, or by regulation of receptor gene expression. Several previous studies have provided evidence of evolutionary arms races at the level of virus-receptor protein interactions. For example, Transferrin Receptor-1 (TfR1) is a key regulator of iron uptake in mammalian cells and is up-regulated when intracellular iron concentrations are low [92]. However, TfR1 is also used for cell entry by viruses such as the Mouse mammary tumor virus (MMTV) and the Machupo virus. Clear evidence of positive selection has been found both on the binding sites of TfR1 for MMTV and Machupo virus and on the corresponding sites in the virus proteins that bind these [101-104]. Mutations at these residues affect receptor-binding interactions and change virulence and host susceptibility, suggesting an ongoing evolutionary arms race. Regulation of

host cell surface receptors can also be an effective defense strategy against virus entry [99, 100, 105, 106]. For example, there appears to be significant variation across human bladder cells for mRNA and protein expression levels of the Coxsackie and Adenovirus Receptor (CAR) gene, another virus-targeted receptor. Thus, the T24 bladder cell line has very low CAR expression and is resistant to virus entry, whereas RT4 cells have high CAR expression level and are highly susceptible to infection [107]. Thus, regulatory changes affecting cell surface receptor levels are related to susceptibility to viral infection. Clearly, however, there may be a tradeoff between reduced receptor expression and the fitness gained by reduced infectivity, which may explain why there are many more published examples of virus-receptor coevolution than for receptor expression evolution (virus-receptor coevolution is also easier to study, so ascertainment bias may also be a factor).

Thus, hosts may adopt different resistance mechanisms at different system levels, e.g., receptor binding vs regulation. However, little previous research has focused on how these different levels of defense mechanisms may evolve in the context of host-pathogen co-evolution. Computational models such as the gene regulatory network evolution model (also known as the Wagner model), that combine a complex genotype-phenotype mapping (describing a gene regulatory network) with evolutionary dynamics have previously been used to address a range of questions concerned with evolution of biological complexity [30, 31]. In previous studies, the gene regulatory network evolution model has been extended to account for different system levels, including transcription factor (TF)-DNA binding interactions [32] and protein-protein interactions (PPI) [33] at the microscopic level, or between two different populations [34] at the macroscopic level. These previous studies [33, 34] showed how robustness and evolvability can evolve to be distributed across different system levels, depending on the model conditions. Here,

we integrate protein-protein interactions (virus-receptor binding) and gene regulatory networks (which control receptor expression) in the context of an evolutionary model that represents both host and pathogen populations.

Viral proteins commonly evolve to mimic receptor binding sites in order to enter host cells through cell surface receptors [96, 101-104]. We introduce a model where the host receptor and the corresponding viral protein are represented as linear sequences and binding is quantified by a similarity score, under the assumption that a close match corresponds to better binding and a higher probability of viral entry. Hosts can evolve to block viral entry either via binding site mismatches or by regulatory changes in receptor protein expression. We further investigate how hosts evolve resistance to different types of viruses: specialists (that target a single receptor) vs generalists (that target many receptors). We consider how the balance between receptor binding and regulation evolves in the context of host-pathogen co-evolution and the need for virus to enter the host cell and the host to block virus entry. More generally, we consider what evolutionary conditions might drive a shift from protein-protein interaction towards gene regulation, and thus increased biological complexity, a key question in the field of evolutionary biology [108, 109]. Furthermore, because we specifically consider host-pathogen coevolution, our study begins to address how complex immune systems may have evolved.

3.2. Model

3.2.1. Host-virus coevolution model

The individual gene regulatory network (GRN) structure and gene expression dynamics largely follows the original gene regulatory network evolution model [17, 110, 111], with 3 primary differences: (i) host individuals are represented by a GRN together with a set of receptor

binding site sequences, (ii) populations follow the dynamics of an SIS model, and (iii) the selection pressure on hosts is given by differential survival probability for the offspring of susceptible vs infected parents and by the rate of disease-related death for infected hosts as selection on the hosts arises from the advantage that resistant offspring have over non-resistant offspring.

A host GRN is represented as a matrix (W) of size $N \times N_{TF}$ where N is the total number of genes, which includes receptor genes (N_R) and the transcription factor genes (N_{TF}) that regulate them. Each element, w_{ij} indicates a regulation of the gene i by a gene product of the gene j, and can represent activation ($w_{ij} > 0$), inhibition ($w_{ij} < 0$), or no regulation ($w_{ij} = 0$). The network density (c) is a parameter of the model and is defined as the fraction of nonzero w_{ij} elements in the matrix W. A founder host individual has a randomly assigned W with a given network density c and with each nonzero w_{ij} element drawn from a Normal distribution, $N(0,1)$. Each row i of the matrix W represents the *cis*-regulatory elements of the i^{th} genes. The GRN is composed of two sub-networks. The first sub-network, from the 1^{st} row to the N_{TF}^{th} row corresponds to the transcription factor (TF) genes and the second sub-network, from the N_{TF+1}^{th} row to the last N^{th} row corresponds to the N_R receptor genes. The expression levels of the N genes at time t are represented as a vector $S(t)$ where the i^{th} element $S_i(t)$ corresponds to the gene expression of i^{th} gene. A sub-vector of $S(t)$ of TF genes ($S_1(t) \sim S_{TF}(t)$) is called $S^{TF}(t)$, and a sub-vector of $S(t)$ of receptor genes ($S_{TF+1}(t) \sim S_N(t)$) is called $S^R(t)$. Initial gene expression $S(0)$ is set as a random binary vector where 0 corresponds to no gene expression and 1 is for full gene expression. Gene expression levels are updated according to the equation $S(t+1) = Sig(W \cdot S^{TF}(t))$, where $Sig(x) = \frac{1}{1+e^{-ax}}$ (a=100) is a sigmoid function which maps

values to gene expression levels in the range $(0,1)$. Here, 0.5 corresponds to basal (unregulated) gene expression. When the gene expression dynamics $S(t)$ reach steady state [31] we simplify gene expression to binary form by applying the function $\varphi(x) = \begin{cases} 0, & x \leq 0.5 \\ 1, & x > 0.5 \end{cases}$, thus defining the phenotype $\hat{S}$.

In the model, we assume there is some degree of functional redundancy for cell surface receptors. Among the total number (N_R) of receptors which can be expressed on the cell surface, a subset (N_{ER}) is required to satisfy the minimum demand for normal host functions. Here we tested $N_{ER} = 1$ or 3 among $N_R = 5$ receptors. For example, $N_{ER} = 1$ indicates that expression of any single receptor is sufficient for host function and any receptor can substitute for any other. At the other extreme, if $N_{ER} = 5$ then all receptors must be expressed and there is no functional redundancy. There are multiple examples showing that different receptors on a host cell can be targeted for virus entry and also that a single host receptor can be targeted by different viruses [90, 91]. Hence, offspring individuals whose phenotypes have fewer expressed receptor genes than N_{ER} ($1 \leq N_{ER} \leq N_R$) are assigned zero fitness since we assume that this is the minimum required for normal host cell functions. The expressed receptor genes produce cell surface receptor proteins that can be targeted by viruses for entry. Each receptor protein is represented as a binary vector of length L, where 0 indicates a polar amino acid and 1 indicates a hydrophobic amino acid. To represent different receptors on the host cell surface, an amino acid sequence is assigned to each receptor protein independently (we avoided having a homogeneous set of initial host receptor proteins as we found this caused population decay due to extremely beneficial conditions for the virus infection). While a host individual is represented with a GRN together with a set of receptor proteins, each virus is represented only by the protein used to enter host cells, represented also as a binary vector of length L.

The initial host population is created in the form of M clones of a founder individual possessing a randomly assigned matrix W and set of receptor amino acid sequences. The host population iterates through cycles of reproduction, mutation and stabilizing selection (similarity to the phenotype of the founder) for 500 time steps in order to generate genetic diversity within the population before the viruses are introduced [31]. Under asexual reproduction each offspring individual is cloned from a random parent, whereas under sexual reproduction each offspring has two random parents and inherits genes (protein sequences and *cis*-regulatory regions) from either parent randomly assuming free recombination among the genes. Since each row represents the *cis*-regulatory region of each gene, sexual reproduction involves copying each row of W from either of the parents for all N genes. GRN mutations change regulatory interactions between genes. As used previously [111], we allow interaction addition ($w_{ij} = 0 \rightarrow w_{ij} \neq 0$), deletion ($w_{ij} \neq 0 \rightarrow w_{ij} = 0$), and modification ($w_{ij} = w'_{ij} \neq 0 \rightarrow w_{ij} = w^*_{ij} \neq w'_{ij}, 0$). The mutation frequency per matrix W is μ including addition (ρ), deletion (ϕ) and modification (δ). ρ and ϕ are set to satisfy $\Delta c = c(t+1) - c(t) = \frac{\mu}{N \cdot N_{TF}} \cdot \{\rho(1 - c(t)) - \phi c(t)\} = 0$ so that the network density (c) remains close to that of the founder. Before contact with viruses, the host population size is fixed and hosts evolve under stabilizing selection to be close to the founder's gene expression phenotype and expressed receptor amino acid sequences. Under stabilizing selection, a host whose phenotype has more than one gene expression difference is not able to survive. Protein mutations involve switching between 0 (polar) and 1 (hydrophobic), where the mutation probability is μ_{hp} per set of receptors. We assume that the amino acid mutations at the virus protein binding site do not affect protein folding. Also for the receptor similarity, we measured a fitness value $f = e^{-\frac{D}{\sigma}}$, where $\sigma = 0.1$ (strong selection) and $D = \frac{\sum_{r \in ER} \sum_{i=1}^{L} |a_{r,i} - a^f_{r,i}|}{|ER| \cdot L}$ (ER: set of

expressed receptors, $|ER|$: the number expressed receptors, $a_{r,i}$: the i^{th} entry of the amino acid sequence of receptor r, $a^{f}_{r,i}$: the i^{th} entry of the amino acid sequence of the founder receptor r), which is the mean L1 distance from the founder amino acid sequence for all expressed receptors.

In preparation for the infection phase, two founder viruses are generated based on protein sequences from host individuals in order to guarantee a high initial transmission rate. Specifically, each founder virus is copied from a receptor protein sequence of a random host, then mutated using the virus protein mutation rate ($\mu_{vp} = 0.1$ per virus protein). Although we tested a case of larger initial virus population size including a greater diversity of founder viruses, we could not find a significant difference from the small initial founder virus population case in terms of the infection strategy of the virus. Hence, in this study, we used two founder viruses for all simulations. Once the host-virus coevolution phase begins, the hosts are divided into susceptible and infected populations and the host population is no longer under stabilizing selection, as hosts need to acquire phenotypic variation to defend against virus entry. Initially all hosts are susceptible and as the founder viruses infect the healthy hosts, those hosts are moved to the infected population. Each individual in the infected group possesses the virus that caused the infection. From this point the population evolves under conditions of co-evolutionary selection and the size of the susceptible (S) and infected (I) groups is allowed to vary. The susceptible and infected population dynamics are inspired by the standard SIS model with births and deaths as shown in the following difference equations:

$$\Delta S = S(t+1) - S(t) = \eta \cdot b \cdot N(t) \cdot \left(1 - \frac{N(t)}{K}\right) - \xi \cdot \frac{r}{N(t)} \cdot S(t) \cdot I(t) - \lambda_N \cdot S(t) + \gamma \cdot I(t) \quad (1)$$

$$\Delta I = I(t+1) - I(t) = \xi \cdot \frac{r}{N(t)} \cdot S(t) \cdot I(t) - (\lambda_N + \lambda_D + \gamma) \cdot I(t) \quad (2)$$

where $N(t) = S(t) + I(t)$, b=growth rate, K=carrying capacity, $\eta = \frac{\#\ of\ survived\ offspring}{\#\ of\ offspring\ candidates}$, r=contact rate, $\xi = \frac{\#\ of\ infections}{\#\ of\ contacts}$ (determined empirically, as described below), $r \cdot \xi$=transmission rate, λ_N=natural death rate, λ_D=disease related death rate, γ=recovery rate. The main difference from the standard ODE SIS model is that ξ and η are determined by the individuals in the population and these parameter values can change as the population evolves. In our model, ξ and η are determined through a complex process that includes random sampling within the population and the evaluation of individual phenotypes. The transmission rate is frequency dependent (i.e., divided by $N(t)$), which assumes that a population occupies an area proportional to its size, i.e., per capita contact rate does not depend on population density, i.e. assuming a wide and unrestricted region affected by infectious viruses [112]. We also use standard assumptions of logistic population growth and that every offspring is initially susceptible. The difference equations dictate the number of offspring that need to be generated, the number of contact events between infected and susceptible hosts, host deaths, and recovered hosts at every time step, but because our model is individual-based, these numeric changes are applied to the actual populations as follows:

The growth term, $\eta \cdot b \cdot N(t) \cdot \left(1 - \frac{N(t)}{K}\right)$, describes the number of offspring, which are generated via sexual or asexual reproduction and mutations in GRN and amino acid sequences are generated as described above. The term $b \cdot N(t) \cdot \left(1 - \frac{N(t)}{K}\right)$ is the total number of offspring candidates who have the stable gene expression and express at least N_{ER} receptors. As candidates who have infected parents are less likely to survive, only a fraction of the candidates (η) can

actually be added to the susceptible population. If phenotypes of the offspring candidates satisfy the criteria of expressing the minimal number (N_{ER}) of receptor genes, and depending on the survival probability, the candidate may be added to the susceptible population. The survival probability is 1 if both parents are susceptible, $k_I < 1$ if both parents are infected, or $\frac{k_I+1}{2}$ if only one parent is infected. Therefore, among the $b \cdot N(t) \cdot \left(1 - \frac{N(t)}{K}\right)$ candidate offspring, only a fraction η of candidates can be added to the susceptible population when k_I is less than 1. Thus, the parameter k_I determines selection due to viral pathogenicity. For the infection term, the number of contacts is $\frac{r}{N(t)} \cdot S(t) \cdot I(t)$. Here, for each contact we choose a random pair of susceptible and infected individuals. We assume that each infected host individual contains a single virus that caused the infection and multiple co-infections were not considered in the model. With each host-virus contact event, the virus mutates the original amino acid sequence at the point of the infection with mutation rate, $\mu_{vp} = 0.1$ per protein. The virus can bind a host receptor if the percentage of one-to-one amino acid pairs that match between the virus and the host receptor exceeds a matching threshold, ϵ_{seqM}. If the virus can bind at least one of the expressed receptors on a susceptible host, then the infection proceeds and the individual moves from the susceptible to the infected population together with the virus that infected it, otherwise the susceptible individual remains in the susceptible population. Successive infection attempts by the same infected individual will involve new mutations with each host-virus contact occurs. Thus, virus transmission will depend on the coevolving host resistance and pathogen virulence. Also, note that the fraction of successful infections ξ in the equations 1 and 2 is determined empirically, rather than as a given parameter.

3.2.2. Parameters

There are parameters at both the level of population dynamics and at the individual level, i.e. governing the regulatory network and the protein sequences (Table 3.1). In this study, we tested a range of parameters including protein binding site amino acid sequence length (L), the minimum number of required expressed receptors (N_{ER}), host protein mutation rate (μ_{hp}), amino acid matching threshold for receptor binding (ϵ_{seqM}), offspring survival probability from both infected parents (k_I) and disease-related death rate (λ_D) to investigate the effect of parameter changes on host resistance evolution. Unless otherwise stated, we used the following parameters: for the population dynamics model, the number of simulations=100, initial host population size M_{init}=150, initial virus population size=2, offspring survival probability from both infected parents k_I=0.8, amino acid matching threshold for receptor binding ϵ_{seqM}=90%, carrying capacity K=1000, growth rate b=0.15, natural death rate λ_N=0.09, disease-related death rate λ_D=0.06, recovery rate γ=0.2, host-virus contact rate r=2. These parameters are chosen to make steady state host population size large enough to investigate evolutionary mechanisms. For the GRN and protein evolution model, virus protein mutation rate $\mu_{vp} = 0.1$, the number of TFs N_{TF}=5, network density c=0.4, mutation rate per W $\mu = 0.1$ with ρ=0.028 and ϕ=0.042 ($\phi + \delta = 1$). Note that $\phi + \delta = 1$, since for an interaction (w_{ij}), deletion and modification are conditional on the interaction being nonzero value ($w_{ij} \neq 0$). These individual level parameters are chosen based on our previous study [34].

Table 3. 1. The list of model parameters at both the level of population dynamics and at the individual level in symbols with descriptions and parameter values used in this study.

Parameter symbol	Description	Values
L	Protein binding site amino acid sequence length	5, 10, 15, 20, 25, 30
μ_{hp}	Host protein mutation rate per a set of	0.002, 0.01, 0.05

	receptors	
μ_{vp}	Virus protein mutation rate	0.1
N_{TF}	The number of transcription factor genes	5
N_R	The number of receptor genes	5
N_{ER}	The minimum number of required expressed receptors	1, 3
ϵ_{seqM}	Amino acid matching threshold for receptor binding	90%, 75%
k_I	Offspring survival probability from both infected parents	0.5, 0.8
ξ	$\frac{\#\ of\ infections}{\#\ of\ contacts}$	Self-determined during simulations
η	$\frac{\#\ of\ survived\ offspring}{\#\ of\ offspring\ candidates}$	Self-determined during simulations
K	Carrying capacity	1000
M_{init}	Initial host population size	150
b	Growth rate	0.15
λ_N	Natural death rate	0.09
λ_D	Disease-related death rate	0.06
γ	Recovery rate	0.2
r	Host-virus contact rate	2
c	Network density	0.4
μ	Mutation rate per gene regulatory network	0.1
ρ	Conditional rate of interaction addition in gene regulatory network	0.028
ϕ	Conditional rate of interaction deletion in gene regulatory network	0. 042
δ	Conditional rate of interaction modification in gene regulatory network	0. 958
σ	Selection pressure	0.1
a	Gene expression mapping sigmoid function parameter	100

3.3. Results

3.3.1. Population dynamics of infection

For many infectious diseases, hosts never achieve long-term immunity due to rapid pathogen divergence. In particular, RNA viruses such as rhinoviruses and coronaviruses mutate so rapidly that even hosts that have recently recovered from an infection can become susceptible again to different strains of the same viruses circulating in the population. The Susceptible-Infectious-Susceptible (SIS) model is a simple infectious disease model that has been widely used to describe population dynamics for rapidly evolving pathogens and their target host populations [113, 114]. We introduce a model of host-virus coevolution that extends the gene regulatory network evolution model of gene regulatory network evolution, integrating it with a discretized form of the SIS model at the population level (see Methods). In our combined model, population sizes can vary, in contrast to the original gene regulatory network evolution model that considered a fixed population size. Since we preserve an explicit representation of each individual genotype in the population, we can observe the evolution of defense and infection mechanisms in both the host and pathogen populations. In its standard form, the SIS model uses fixed values to describe parameters such as the infection transmission rate. However, on evolutionary timescales, parameters such as host susceptibility and pathogen virulence are likely to vary over time and consequently key model parameters such as the transmissibility, ξ, will also change. In our model, each host genotype is represented explicitly with a gene regulatory network and the corresponding receptor protein sequences (Figure 3.1). Each virus is represented explicitly with a receptor binding protein sequence, that will be compared to the host receptor sequences during contact (attempted infection) events (Figure 3.1). Hence, rather than determining the rate of infection based on a fixed parameter, as in the standard SIS model, we allow the contacting host and pathogen phenotypes to determine infection events. Specifically,

the key transmission parameter ($\xi = \frac{\#\ of\ infections}{\#\ of\ contacts}$) that determines the infection rate ($r \cdot \xi$) changes as both hosts and viruses evolve. Analytically, the steady state susceptible and infectious population sizes are $\tilde{S} = \frac{\delta_I}{r \cdot \xi} \cdot K \cdot \left\{1 - \frac{1}{b \cdot \eta} \cdot \left(\lambda_N + \lambda_D\left(1 - \frac{\delta_I}{r \cdot \xi}\right)\right)\right\}$ and $\tilde{I} = \left(1 - \frac{\delta_I}{r \cdot \xi}\right) \cdot K \cdot \left\{1 - \frac{1}{b \cdot \eta} \cdot \left(\lambda_N + \lambda_D\left(1 - \frac{\delta_I}{r \cdot \xi}\right)\right)\right\}$ respectively when $r \cdot \xi \neq 0$ and $\frac{b \cdot \eta - \lambda_N}{\lambda_D} > 1 - \frac{\delta_I}{r \cdot \xi} > 0$ where $\delta_I = \lambda_N + \lambda_D + \gamma$. Different steady state values of ξ lead to different $\tilde{S}$ and $\tilde{I}$ since these population sizes ultimately depend on the value of ξ. Since our main interest is the evolution of host resistance mechanisms, we only analyzed cases where the mean population size over time is greater than the initial susceptible population size (M_{init}=150). In cases where the mean total population size < M_{init} (Figure 3.2), we found that the susceptible population was too small to investigate and these cases mostly occur when the extremely infectious viruses appear which can spread widely and make the host population sick.

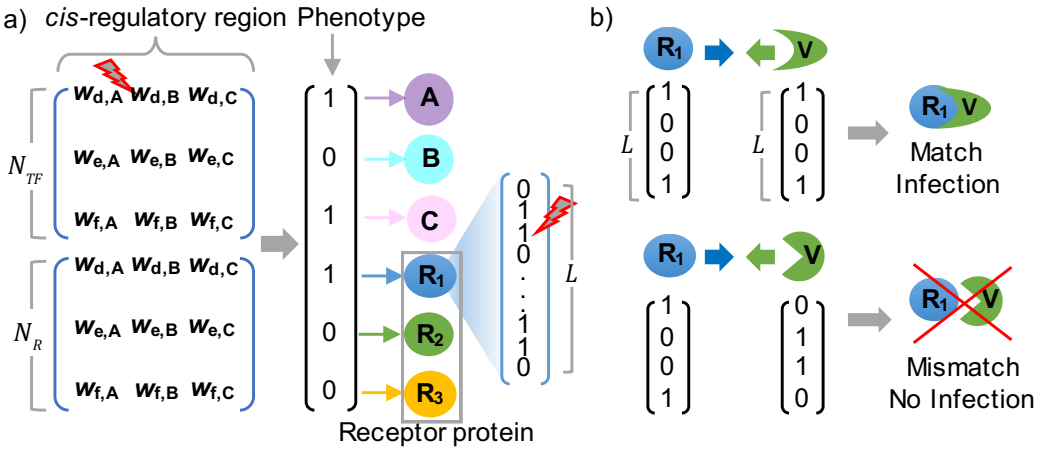


Figure 3. 1. Diagram of gene regulatory network (GRN) and host-virus interaction scheme. a) the GRN is composed of a transcription factor regulation sub-network and a receptor protein coding regulation sub-network. Mutations at the network level can be used to shut down the targetable receptor. Mutations at the protein level can result in a protein mismatch to block virus protein binding. b) If more than ϵ_{seqM}% of amino acids are one-to-one matched, we assume the virus protein can bind to the matched receptor (top). If less than the threshold (ϵ_{seqM}) are matched, we assume the virus protein fails to bind the receptor.

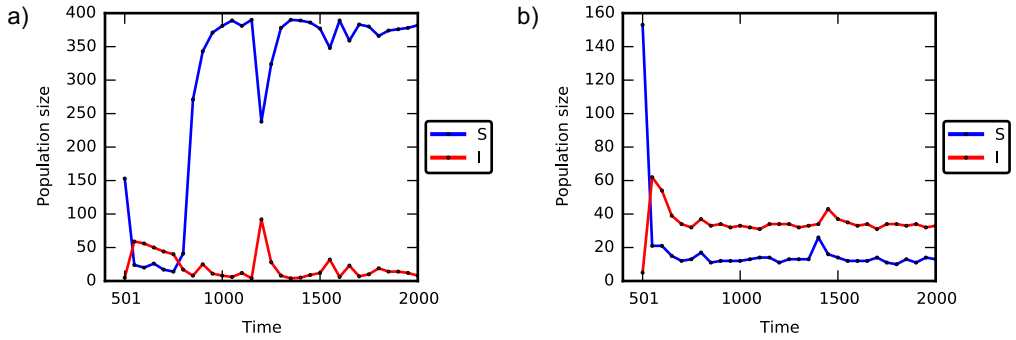


Figure 3. 2. Two different types of susceptible and infectious population dynamics. Typical population dynamics of a) healthy population case where the mean host population size is greater than the initial host population size and b) sick population case where the mean host population size is less than the initial population and the population is composed of more infected hosts than healthy hosts. (L=10, $\boldsymbol{N_{ER}}$=3, $\boldsymbol{\mu_{hp}}$=0.002, $\boldsymbol{\epsilon_{seqM}}$=75%, $\boldsymbol{k_I}$=0.8)

We measured the steady state transmissibility (ξ), defined here as the mean value of ξ across the last 250 time points in each simulation, and considered how this measure changed under different conditions such as the protein binding sequence complexity (length, L), host protein mutation rate (μ_{hp}), the number of required expressed receptors (N_{ER}), the threshold above which the virus and receptor proteins are considered to have matched (ϵ_{seqM}), the survival rate from infected parents (k_I) and the disease-related death rate (λ_D). As shown in Figure 3.3, higher receptor binding sequence complexity (L) and higher host protein mutation rates (μ_{hp}) tend to generate lower transmissibility ξ and are therefore disadvantageous to virus transmission. Similarly, when more receptors have to be expressed on the host cell surface (higher N_{ER}), there are more ways in which viruses can attempt receptor binding and consequently, ξ tends to increase together with the number of required expressed receptor (N_{ER}), at least when the receptor binding complexity is low (Figure 3.4a). For similar reasons, the transmissibility ξ also

increases for lower matching threshold (ϵ_{seqM}) value, such that when protein binding sequence complexity (L) is low, reducing the matching threshold (ϵ_{seqM}) dramatically increases virus transmission whereas for complex receptor binding, it does not have an advantageous effect on ξ (Figure 3.4b). That transmissibility ξ increases only in the case of low complexity binding can be explained by the way viruses target host receptors, as explained in the next section. Intuitively, when a survival rate from infected parents (k_I) is low, non-resistant offspring have much lower fitness (if infected) than resistant offspring, and thus resistant individuals should increase in frequency. This would actually tend to decrease ξ which is the opposite of what we observe. However, we found that in practice, it is more common for a low k_I value to cause population decay and a large decrease in the number of contacts between host and virus individuals as shown in (Figure 3.5). A reduced number of contacts causes a larger decrease in the denominator of ξ ($\frac{\#\ of\ infections}{\#\ of\ contacts}$), and therefore leads to a net increase in ξ (Figure 3.4c). The observation of higher ξ as a consequence of a high disease related death rate (λ_D) is due to the same reason as for low k_I (Figure 3.4d). In sum, the virus transmissibility is dependent on various conditions for different underlying reasons. We now consider in greater detail why and how these variables affect the host and virus population dynamics and virus transmission.

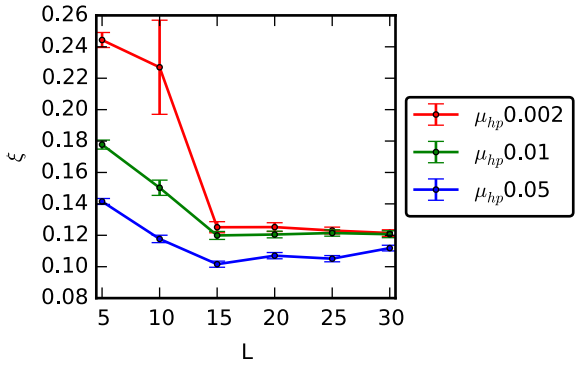


Figure 3. 3. Transmissibility changes for different receptor binding complexity and host protein mutation rate. The mean transmissibility (***ξ***) for the last 250 time points (Error bar: one std. dev. over 100 simulations). ***ξ*** increases as the receptor binding complexity decreases (shorter ***L***) in which case viruses can target multiple receptors and as the host protein mutation rate (***μ_{hp}***) decreases which is due to the more limited speed of protein mutations to counteract the rapidly evolving viruses.

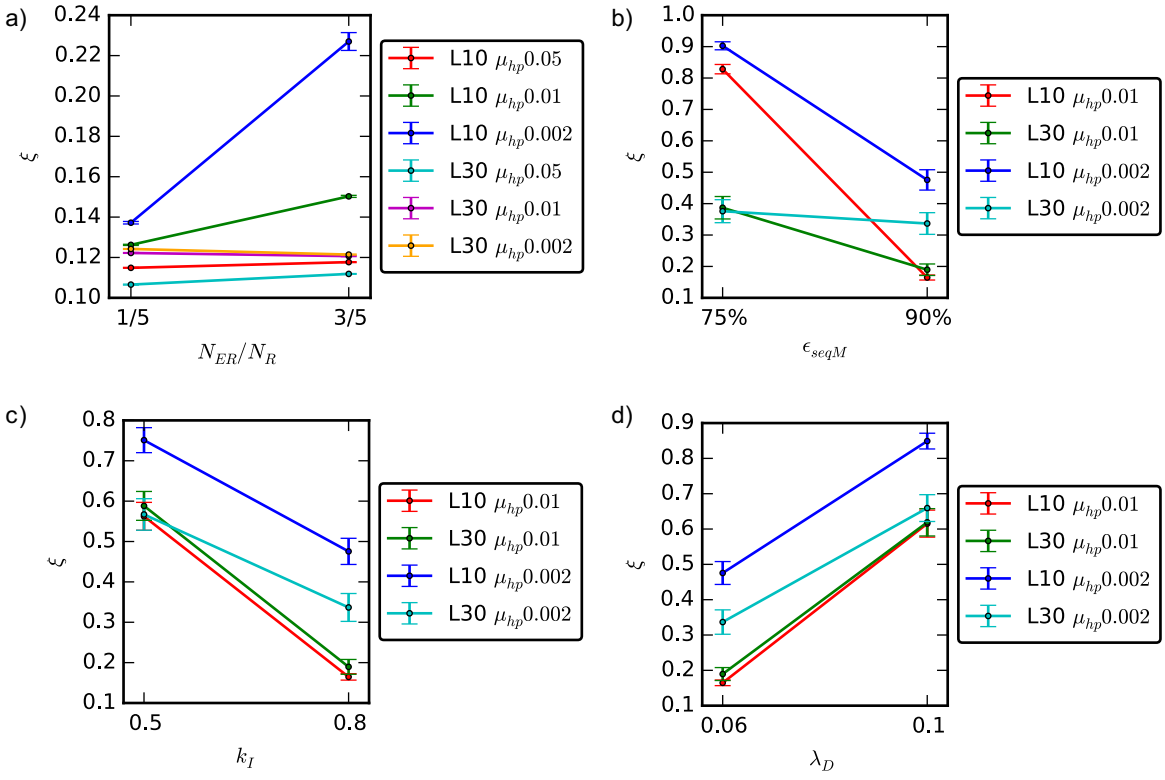


Figure 3. 4. Transmissibility changes for different conditions. The mean transmissibility (***ξ***) for the last 250 time points (Error bar: one std. dev. over 100 simulations). a) ***ξ*** increases as the number of required receptor expression (***N_{ER}***) increases when the binding complexity (***L***) is low.

For low receptor binding threshold ($\boldsymbol{\epsilon_{seqM}}$), low survival rate from both infected parents ($\boldsymbol{k_I}$) and high disease related death rate ($\boldsymbol{\lambda_D}$), population dynamics generally follows that shown in Figure 3.2b. Hence, in b), c) and d) we considered all 100 simulations for the comparison of mean $\boldsymbol{\xi}$ values. $\boldsymbol{\xi}$ increases as (b) the receptor binding site matching threshold ($\boldsymbol{\epsilon_{seqM}}$) decreases, as (c) the survival rate from both infected parents ($\boldsymbol{k_I}$) decreases and as (d) disease related death rate ($\boldsymbol{\lambda_D}$) increases.

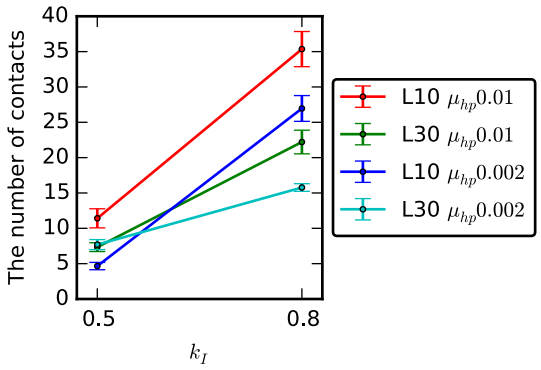


Figure 3. 5. The number of contacts between host and parasite populations for different offspring survival rate from infected parents. The number of contacts between host and parasite populations decreases when offspring survival rate from infected parents ($\boldsymbol{k_I}$) is low (Error bar: one std. dev. over 100 simulations).

3.3.2. Host resistance strategy depends on the number of targeted receptors

Since receptor-virus protein binding enables virus entry and determines whether the infection succeeds, the virus's ability to target multiple receptors and host's ability to escape virus protein binding will have a significant impact on host resistance and viral pathogenicity. Hence we measured the number of targeted receptors across a variety of different conditions. We next show how the number of targeted receptors can change depending on the receptor binding complexity (protein sequence length, L), the number of required expressed receptors (N_{ER}), protein binding threshold (ϵ_{seqM}), the survival rate from infected parents (k_I) and the disease-

related death rate (λ_D). As each simulation proceeded, we measured the frequency with which multiple receptors are targeted simultaneously and also used the Gini coefficient to measure the unevenness in the distribution of targeted receptors among the newly infected hosts throughout the simulation (see Methods). Thus, for example, when the frequency of multi-receptor matching is low, this indicates that mostly a single receptor is being targeted by the virus. However, this does not guarantee that the virus population targets the same specific receptor or whether different subpopulations are targeting distinct receptors. In this case, when the Gini coefficient of targeted receptors is high, this indicates that all viruses target a common receptor and when the Gini coefficient is low, this implies that the matched receptor for each host is different and that viruses have diversified into subpopulations by targeting different receptors.

When binding complexity (***L***) is low, viruses can target different receptors by means of a few amino acid mutations, whereas when receptor binding complexity is high, targeting multiple receptors is more difficult since the different receptors are likely separated by more mutations. Hence, as shown in (Figure 3.6), when ***L*** is short, multiple receptors are often targeted simultaneously and the frequency of each receptor being targeted is not highly variable (low Gini coefficient). Considering this, more permissive receptor binding (lower $\boldsymbol{\epsilon_{seqM}}$), increases the chances for multiple receptor targeting when ***L*** is short (Figure 3.7c, d). On the other hand, when binding complexity is high, a single receptor is usually targeted and the Gini coefficient is close to 1 indicating there are usually one or two dominant targeted receptors (Figure 3.6). Furthermore, in this case, reducing the receptor binding threshold does not help viruses target multiple receptors (Figure 3.7c, d). These results indicate that for complex receptor binding, one or two receptors are targeted for virus entry and that there is no switch from one targeted receptor to another (Figure 3.6). Based on this observation, as expression of more distinct

receptors is required (higher $\boldsymbol{N_{ER}}$), multiple receptors can be targeted and at the same time the Gini coefficient decreases only when receptor binding complexity is low (short $\boldsymbol{L}$). On the other hand, when receptor binding is complex (long $\boldsymbol{L}$), increasing $\boldsymbol{N_{ER}}$ does not allow more receptors to be targeted by viruses (Figure 3.7a, b). Hence the number of required expressed receptors only impacts the strategy of the virus when the receptor binding is less complex (short $\boldsymbol{L}$).

Interestingly, the survival rate of offspring from infected parents also affects how the viruses target receptors. As we explained in the previous section, a low survival rate from infected parents ($\boldsymbol{k_I}$) causes the host population to become sick (the mean host population size is less than the initial population and the population is composed of more infected hosts than healthy hosts) and thus the population size decays. Consequently, as shown in Figure 3.13d, e and f, we observe that variation within the host population decreases, suggesting that viruses will need to specialize on binding to specific receptors (Figure 3.7e, f). Specific receptor targeting as a consequence of high disease related death rate ($\boldsymbol{\lambda_D}$) arises for the same reason as for low $\boldsymbol{k_I}$ (Figure 3.7g, h). We tested the effect of diversity in the initial virus population on the number of targeted host proteins. We compared a case with a highly diverse initial virus population to the default case of two initial viruses. Thus, given an initial population of 15 distinct founder viruses, each three viruses were chosen to bind a distinct host receptor. With $\boldsymbol{L = 30, \mu_{hp} = 0.002}$ and $\boldsymbol{N_R = 5}$, all virus strains except one went extinct. In this case, the frequency of multi-receptor targeting was 0.04±0.04 and unevenness of targeting receptors (Gini coefficient) was 0.793±0.009 which is close to the values for the 2 founder virus case. Even with $\boldsymbol{L = 10, \mu_{hp} = 0.002}$ and $\boldsymbol{N_R = 5}$, we could not find a significant difference from the 2 founder case. Here, the frequency of multi-receptor targeting was 0.16±0.14 and unevenness of targeting receptors (Gini coefficient) was 0.70±0.08. In sum, receptor binding complexity ($\boldsymbol{L}$) affects viruses by determining the variety of

targetable receptors, although this also is dependent on parameters such as $\boldsymbol{N_{ER}}$ and $\boldsymbol{\epsilon_{seqM}}$. Also indirect causality between host population diversity and parameters, $\boldsymbol{k_I}$ and $\boldsymbol{\lambda_D}$ has an influence on the specificity of targetable receptors. So far, we considered how viruses behave and choose infection strategies for different conditions. We next explore how hosts react to virus infection strategies differently depending on the various environments.

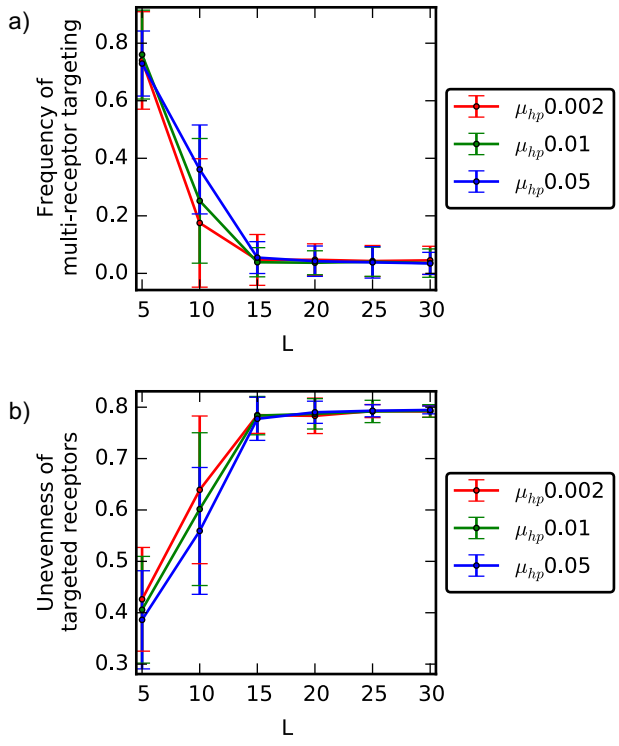


Figure 3. 6. Two different virus infection strategies: Targeting a specific receptor or non-specific multiple receptors. a) The fraction of time points that multiple receptors are targeted simultaneously and b) the Gini coefficient of the frequency of targeted receptors for different receptor binding complexities (***L***s) (Error bar: std. dev. over 100 simulations). A lower Gini coefficient (close to zero) indicates evenness and one that is close to one indicates inequality. As the receptor binding complexity increases (longer ***L***) viruses target a specific receptor and do not change the target receptor over time.

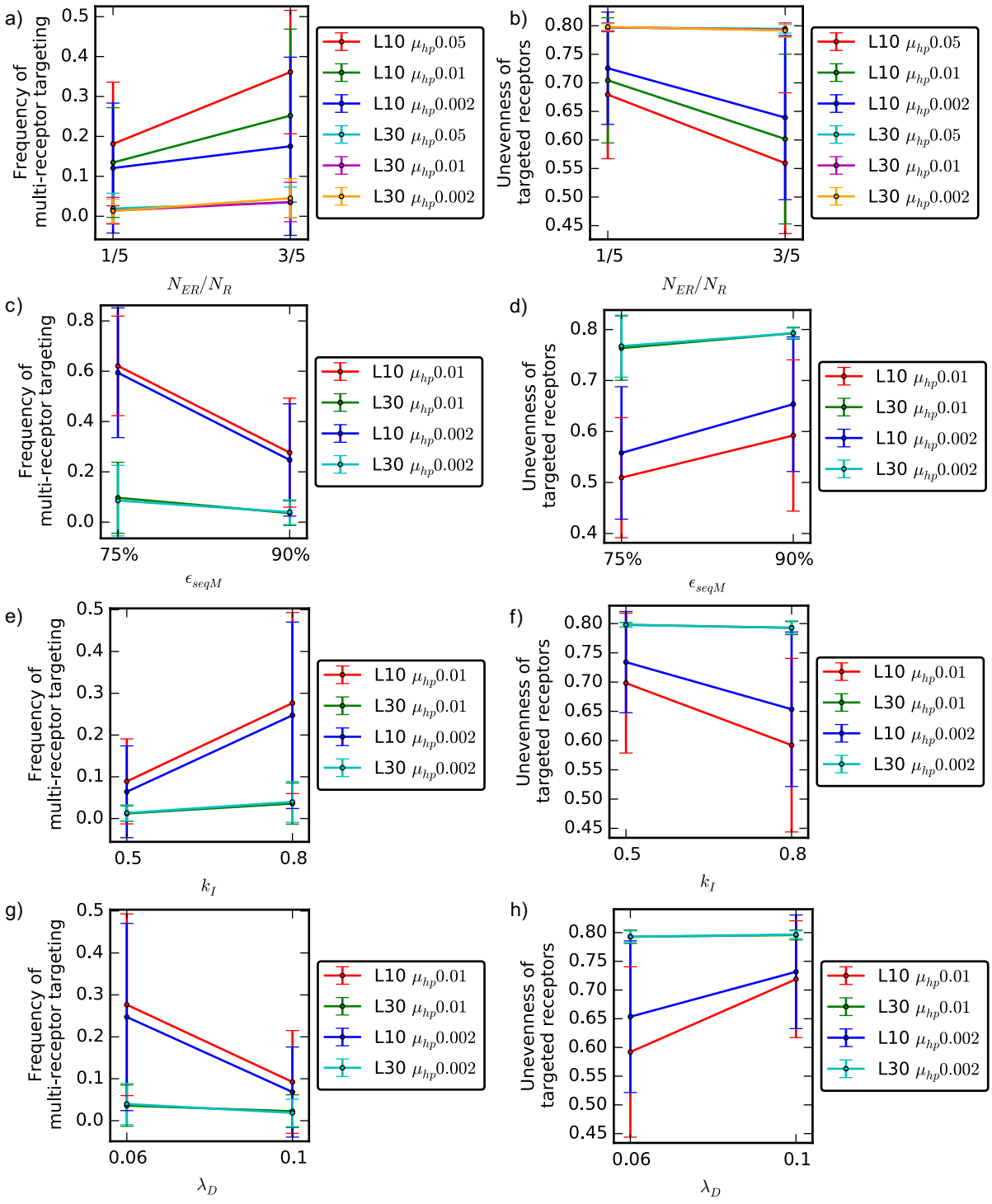


Figure 3. 7. Viruses change their receptor targeting strategy under different conditions. The first column is the fraction of time points that multiple receptors are targeted simultaneously and the second column is the Gini coefficient of the frequency of targeted receptors (Error bar: one std. dev. over 100 simulations). a, b) When the binding complexity is low, a greater required number of expressed receptors ($\boldsymbol{N_{ER}}$) causes viruses to target multiple receptors simultaneously.

However, when the binding complexity is high, a higher required number of expressed receptors does not change the targeting to a multiple receptor binding strategy. For low receptor binding threshold ($\boldsymbol{\epsilon_{seqM}}$) and survival rate from both infected parents ($\boldsymbol{k_I}$), population dynamics generally follows the trend shown in Figure 3.2b. Hence, in c~h) we considered all 100 simulations for the comparison of the fraction of time points that multiple receptors are targeted simultaneously and the Gini coefficient of the frequency of targeted receptors. c, d) The low amino acid matching threshold for the receptor binding ($\boldsymbol{\epsilon_{seqM}}$) facilitates viruses to target multiple receptors. e, f) The low survival rate of an offspring from both infected parents results in viruses targeting more specific receptors for more robust receptor binding. g, h) The high disease related death rate ($\boldsymbol{\lambda_D}$) causes more specialized receptor targeting.

3.3.3. Evolved preference for resistance using network rewiring

Hosts can adopt two different resistance strategies in the model: 1) Gene regulatory network rewiring to switch a targeted receptor off and 2) protein binding site changes to block protein binding to a targetable receptor. Here we consider how hosts balance the usage of these two strategies and what conditions determine their relative preference. At each time step the most frequently targeted receptor is identified among the set of newly infected hosts and from here we measure how often successful resistance events use network rewiring to shut down the most targetable receptor rather than protein sequence changes. We proceed by counting the fraction of hosts who resisted successfully and that do not express the most frequently targeted receptor. If there are multiple equally frequent most targeted receptors, we use the mean frequency across those receptors. The fraction of resisted hosts using network rewiring was measured at every time point. We then accumulated these measurements over all time points throughout the simulation and if the overall use of network rewiring resistance was higher than protein level resistance, we counted the simulation as preferential to rewiring. We subsequently measured the fraction of simulations for which this occurred to quantify the relative use of

rewiring across many simulations. Using this measure, we find that GRN rewiring is preferentially used as protein binding complexity increases (Figure 3.8). This outcome relates to the number of targeted receptors since when protein binding is more complex, the virus most often targets a single receptor and therefore down-regulating the targetable receptor is usually an effective strategy. Conversely when protein binding is low complexity, viruses are able to enter the host cell by binding multiple receptors and therefore rewiring is a less effective host strategy for resistance. As the host protein mutation rate ($\boldsymbol{\mu_{hp}}$) decreases, hosts also use GRN rewiring more often due to the reduced ability to catch up with the relatively fast-evolving virus proteins (Figure 3.8). As we increase the number of receptors that need to be expressed ($\boldsymbol{N_{ER}}$) then combinatorially there are fewer possible phenotypes for a given number of required receptors, and viruses have more chances to bind to the different receptors so that the frequency of resistance using GRN rewiring decreases (Figure 3.9a). Reducing the protein matching threshold also favors the protein interaction level (Figure 3.9b). Lastly, at low survival rate ($\boldsymbol{k_I}$) from infected parents and at high disease related death rate ($\boldsymbol{\lambda_D}$), viruses tend to target more specific receptors, which is due to population size decay and low population diversity (Figure 3.7e~h). In fact, as shown in (Figure 3.10g, i), the *potential* for resistance (which will be explained in the following paragraph) via network rewiring increases. However, the small population size and low variation do not allow this potential to be realized. This explains the apparently contradictory result of (Figure 3.9c, d), where the observed (as opposed to potential) number of resistance events occurring via GRN decreases when $\boldsymbol{k_I}$ is low but also when $\boldsymbol{\lambda_D}$ is high. Hence, unlike with $\boldsymbol{L}$, $\boldsymbol{N_{ER}}$ and $\boldsymbol{\epsilon_{seqM}}$, we observed that low $\boldsymbol{k_I}$ and high $\boldsymbol{\lambda_D}$ did not promote resistance via network rewiring (Figure 3.9c, d). In sum, hosts choose a resistance mechanism depending on the virus infection strategy and their defense ability relative to viruses (how fast they react to

the fast evolving viruses). In the next section, we consider the temporal dynamics of hosts with respect to regulatory network and receptor protein binding evolution.

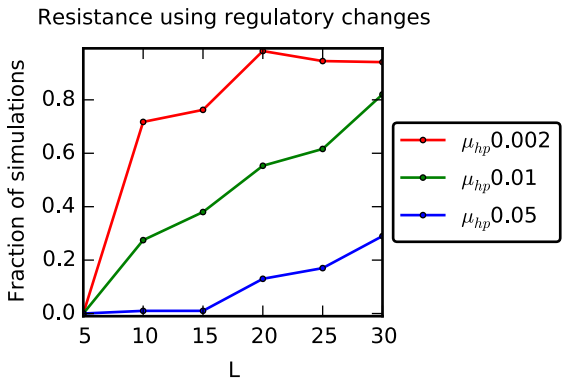


Figure 3. 8. Preference for resistance using gene regulatory network (GRN) rewiring rather than protein mutations. The fraction of simulations where GRN rewiring strategy is used more often than protein binding site change for successful resistance under different protein binding complexities (***L***s) and host receptor sequence mutation rates ($\boldsymbol{\mu_{hp}}$). In a more complex receptor binding system, hosts tend to select the GRN rewiring strategy more often than the protein mutation strategy due to the single receptor targeting infection strategy. Since low $\boldsymbol{\mu_{hp}}$ means a lower rate of protein mutations to counteract the rapidly evolving viruses, hosts tend to favor a protein mutation strategy less.

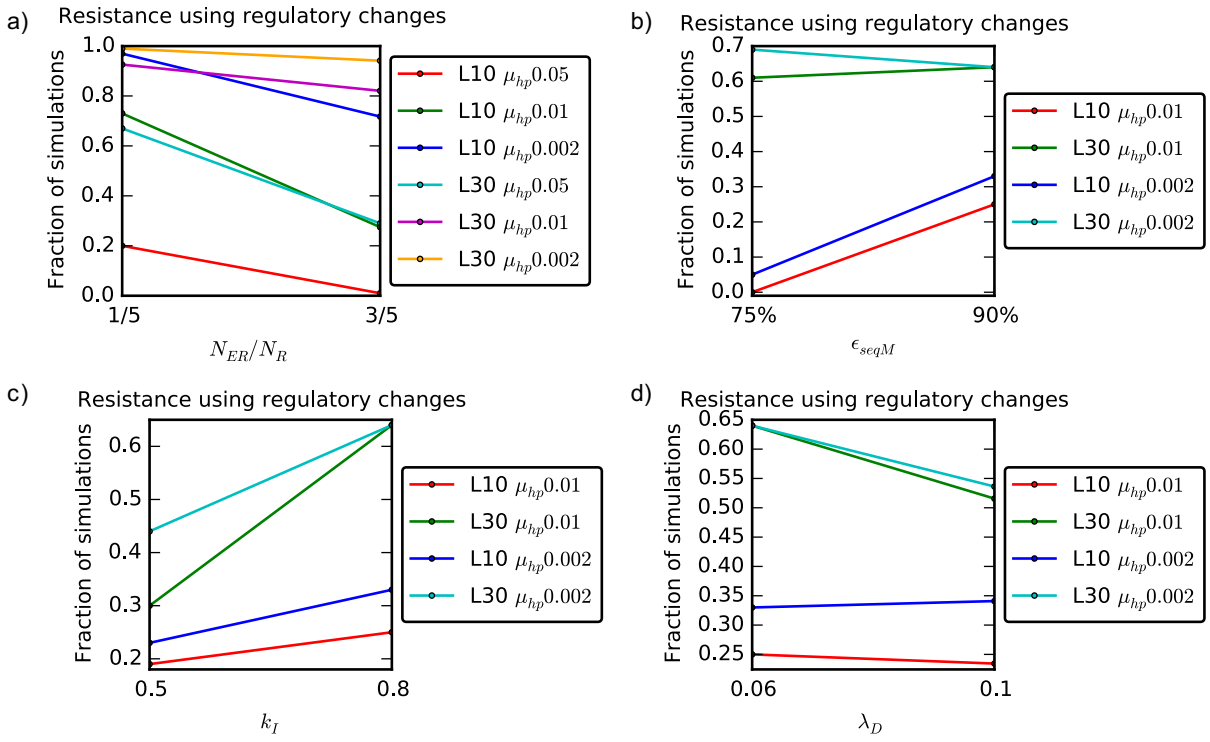


Figure 3. 9. Preference for resistance using gene regulatory network (GRN) rewiring to protein mutations under different conditions. The fraction of simulations where GRN rewiring strategy is used more often than the protein binding site change strategy for resistance for different a) required number of expressed receptors ($\boldsymbol{N_{ER}}$), b) amino acid matching threshold for the receptor binding ($\boldsymbol{\epsilon_{seqM}}$), c) survival rate from both infected parents ($\boldsymbol{k_I}$) and d) disease related death rate ($\boldsymbol{\lambda_D}$). For low $\boldsymbol{\epsilon_{seqM}}$, $\boldsymbol{k_I}$ and $\boldsymbol{\lambda_D}$, the population dynamics generally follows that shown in Figure 3.2b. Hence, in b, c, d) we considered all 100 simulations for the comparison of the preference for resistance using GRN rewiring to protein mutations. a) As more receptors are required to be expressed (higher $\boldsymbol{N_{ER}}$), hosts preferentially use GRN rewiring less often than protein mutations. b) When the binding complexity is low, for lower amino acid matching threshold for the receptor binding ($\boldsymbol{\epsilon_{seqM}}$), hosts do not preferentially select GRN rewiring strategy. c) When $\boldsymbol{k_I}$ is low, hosts do not favor the GRN rewiring strategy. d) When the disease related death rate ($\boldsymbol{\lambda_D}$) is high, hosts less favor the GRN rewiring strategy for resistance.

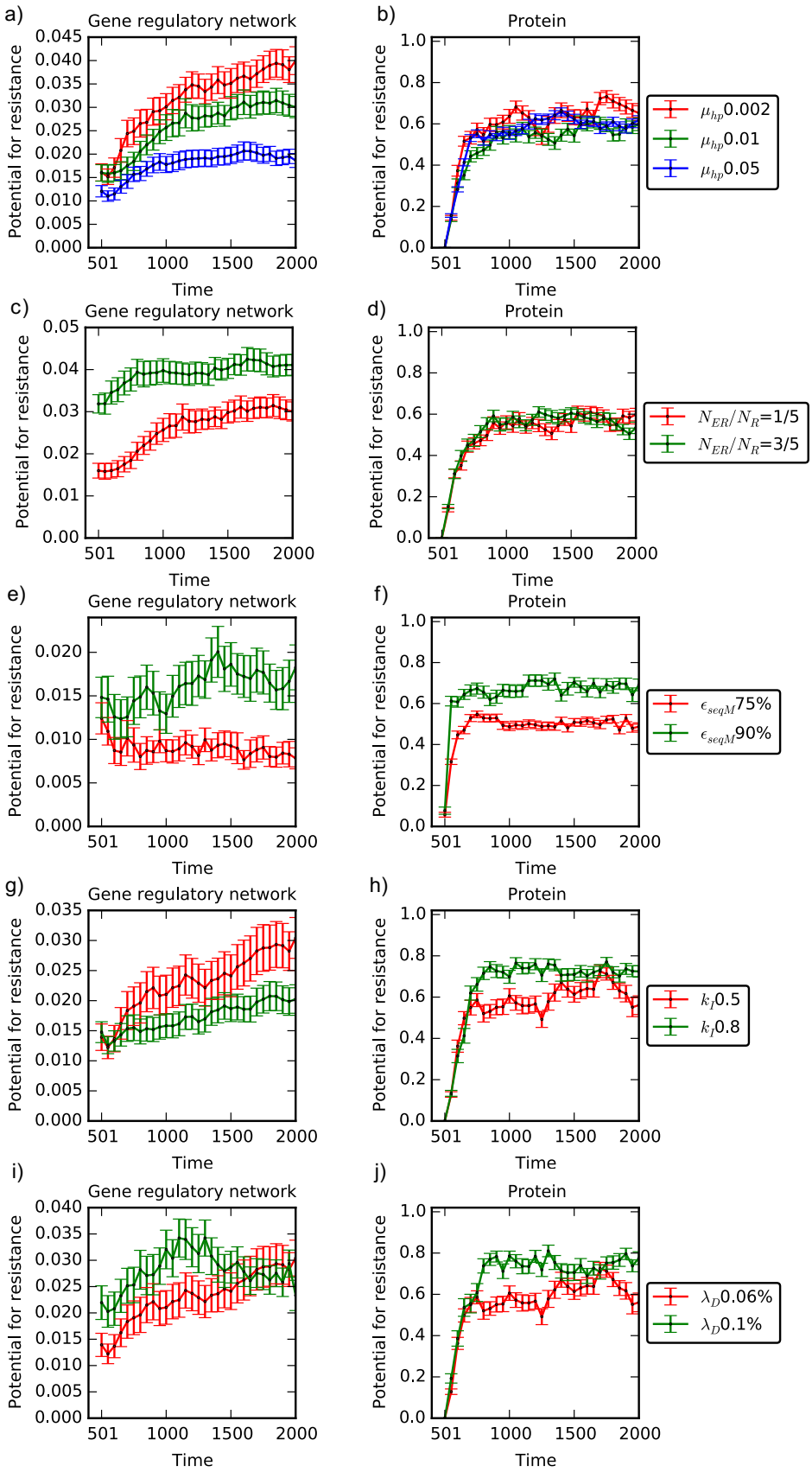

a)
Gene regulatory network
Potential for resistance
Time
b)
Protein
Potential for resistance
Time
μ_{hp}0.002
μ_{hp}0.01
μ_{hp}0.05
c)
Gene regulatory network
Potential for resistance
Time
d)
Protein
Potential for resistance
Time
N_{ER}/N_R=1/5
N_{ER}/N_R=3/5
e)
Gene regulatory network
Potential for resistance
Time
f)
Protein
Potential for resistance
Time
ϵ_{seqM}75%
ϵ_{seqM}90%
g)
Gene regulatory network
Potential for resistance
Time
h)
Protein
Potential for resistance
Time
k_I0.5
k_I0.8
i)
Gene regulatory network
Potential for resistance
Time
j)
Protein
Potential for resistance
Time
λ_D0.06%
λ_D0.1%

Figure 3. 10. Evolutionary potential for resistance in the gene regulatory network and receptor proteins for different conditions. For susceptible host population, the ability to resist using GRN rewiring ($\mathbf{1^{st}}$ column) and protein binding site changes ($\mathbf{2^{nd}}$ column) is measured for different a, b) host protein mutation rates ($\boldsymbol{\mu_{hp}}$), c, d) number of required expressed receptors ($\boldsymbol{N_{ER}}$), e, f) amino acid matching threshold for the receptor binding ($\boldsymbol{\epsilon_{seqM}}$), g, h) survival rate from both infected parents ($\boldsymbol{k_I}$) and i, j) disease related death rate ($\boldsymbol{\lambda_D}$) (Error bar: std. dev. over 100 simulations). For low $\boldsymbol{\epsilon_{seqM}}$ and $\boldsymbol{k_I}$, population dynamics generally follows that of Figure 3.2b. Hence, in e~h) we considered all 100 simulations for the comparison of the resistance potentials. a, b) For lower $\boldsymbol{\mu_{hp}}$, hosts evolve a GRN based strategy ($\boldsymbol{L}$=30, $\boldsymbol{\mu_{hp}}$=0.01, $\boldsymbol{\epsilon_{seqM}}$=90%, $\boldsymbol{k_I}$=0.8). c, d) When expression of more receptors is required, hosts evolve the potential for resistance using GRN rewiring to higher level. ($\boldsymbol{L}$=30, $\boldsymbol{N_{ER}}/\boldsymbol{N_R}$=3/5, $\boldsymbol{\epsilon_{seqM}}$=90%, $\boldsymbol{k_I}$=0.8), e, f) When receptor binding is simple (short $\boldsymbol{L}$), for reduced $\boldsymbol{\epsilon_{seqM}}$ hosts does not necessarily evolve the potential for a GRN rewiring strategy ($\boldsymbol{L}$=10, $\boldsymbol{\mu_{hp}}$=0.002, $\boldsymbol{N_{ER}}/\boldsymbol{N_R}$=3/5, $\boldsymbol{k_I}$=0.8). g, h) Selection pressure triggered by the low $\boldsymbol{k_I}$ evolves the potential for GRN rewiring strategy ($\boldsymbol{L}$=30, $\boldsymbol{\mu_{hp}}$=0.002, $\boldsymbol{N_{ER}}/\boldsymbol{N_R}$=3/5, $\boldsymbol{\epsilon_{seqM}}$=90%). i, j) The potential for resistance using network rewiring increases both for low and high diseases related death rates ($\boldsymbol{\lambda_D}$).

3.3.4. Evolutionarily gained potential to switch from infectious to resistance using GRN rewiring and protein mutations

In the previous section, we showed that hosts determine the resistance strategy between GRN rewiring and protein binding site mutation depending on factors such as binding site complexity and mutation rate relative to that of the virus. We now consider the evolution of the potential within the population to resist future virus contact events. For each virus in the infected group, we selected all susceptible hosts in the population that can be potentially infected by that virus and measure how efficiently each host can avoid infection via a random mutation either in its GRN or in protein binding sites. Every regulatory interaction in the GRN was mutated multiple times and we then measured how often it switched to becoming resistant as a consequence of these network perturbations. Similarly, for each matched receptor, we mutate the receptor using the host protein mutation rate at each site (as would occur during the simulation)

and measured the average fraction of such perturbations that caused a switch to resistance. The reason for using the same protein mutation rate that is used within the simulation rather than a single random amino acid mutation for the perturbation is that the impact of a single site amino acid mutation differs depending on the protein binding site length (L). For example, when L is long, a chance of switching from infectious to resistible is very low, whereas when L is short, a host can easily switch from infectious to resistible.

For resistance acquired via regulatory rewiring, the ability to resist increases only when the protein complexity is high (Figure 3.11a blue and green lines), while it does not increase when the protein binding complexity is low (red line). It is plausible that when the protein binding complexity is low, since network rewiring is not a good resistance strategy (Figure 3.8) due to multiple receptor binding site matches by viruses (Figure 3.6), it is unnecessary for individuals to evolve network rewiring potential and for this reason few perturbations are expected to change receptor gene expression to switch the targetable receptor off. In contrast, when the protein binding complexity is high so that the targeted receptor is specialized to one receptor (Figure 3.6) and switching targetable receptor off by network rewiring is adopted by hosts (Figure 3.8), hosts evolve the potential to resist by network rewiring. In contrast, for resistance via protein mutations, we observed that under all conditions hosts rapidly evolve the ability to acquire resistance via protein binding site changes (Figure 3.10 and Figure 3.11b) because the protein binding site mutations can directly affect virus protein binding.

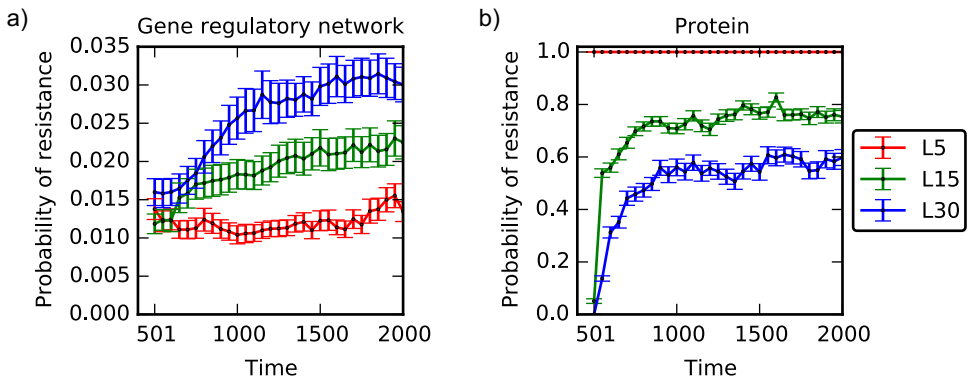


Figure 3. 11. Trade-offs in the resistance potential between the gene regulatory network and receptor proteins. For the susceptible host population, the ability to resist using a) GRN rewiring and b) protein binding site changes is measured for different receptor binding complexities (Error bar: std. dev. over 100 simulations). As the receptor binding complexity increases, hosts increase evolutionary potential more on the GRN while decreasing it on receptor proteins ($\boldsymbol{\mu_{hp}}$=0.01, $\boldsymbol{N_{ER}/N_R}$=3/5, $\boldsymbol{\epsilon_{seqM}}$=90%, $\boldsymbol{k_I}$=0.8).

We also observed that there is an apparent tradeoff in that, as the resistance ability via rewiring increases (Figure 3.11a) with receptor binding complexity, the ability to resist using binding site mutations decreases (compare order of curves in Figure 3.11a vs Figure 3.11b). The complexity of the protein-protein interaction appears therefore to be an important factor driving the transition toward resistance using regulation and thus leading to higher GRN complexity. As expected, when the protein mutation rate is low, hosts will use GRN rewiring more for resistance as a consequence of the limited capacity for protein mutations to coevolve with the viruses (Figure 3.10a, b). The ability to resist using network rewiring also depends on the number of required expressed receptors (N_{ER}). As more receptors are required to be expressed (N_{ER}), viruses have a greater probability of targeting more than one receptor. Hence, as shown above in (Figure 3.9a), the fraction of simulations where GRN rewiring is used in preference to protein mutation decreases for higher values of N_{ER}. However, for the same reason, hosts are under pressure to evolve the ability to resist using network rewiring more when more receptors are

required to be expressed (Figure 3.10c, d). In the (Figure 3.7c, d), in higher matching threshold (ϵ_{seqM}) condition, viruses are not able to target multiple receptors and the fraction of simulations where GRN rewiring is preferentially used also increases (Figure 3.9b). Consequently, high ϵ_{seqM} results in evolution of the potential to resist infection using GRN (Figure 3.10e, f). A lower survival rate from infected parents induces viruses to target specific receptors (Figure 3.7e, f). Therefore, for such viruses, hosts are evolved to increase the ability to resist using GRN rewiring to shut down the targetable receptor (Figure 3.10g, h).

So far, we explored various conditions that can promote the evolution of the ability to resist using GRN rewiring. Interestingly, receptor binding complexity balances the usages of GRN rewiring vs amino acid mutations for resistance. Resistance via protein binding site mutation is much higher than that using network rewiring under all conditions. This may explain why receptor binding site mutations have been reported often for virus entry defense mechanisms in contrast to resistance via regulatory changes.

3.3.5. Genetic diversity and host range

In many previous studies it has been shown that antagonistic coevolution between host and pathogen populations correlates with increased genetic diversity [88, 115]. We checked that the diversity of the regulatory network, the phenotype and the protein sequence all increase throughout the coevolution phase (Figure 3.12). To quantify diversity we used the Margalef index [116], an ecological measure of biodiversity that takes into account the expected increase in species sampled as a consequence of increased sample size ($\frac{\text{the number of genetic variants}-1}{\ln(\text{total number of individuals})}$). After we simplified each GRN using the sign of each interaction matrix entry (e.g., -0.8 to -1 and

+0.8 to 1), we measured the GRN diversity of a susceptible host group as $\frac{\text{the number of distinct GRNs}-1}{\ln(\text{susceptible individuals})}$. We found that diversity of GRNs, phenotypes and receptor protein sequences all increased throughout the coevolutionary phase, showing that coevolution between hosts and viruses is an important factor in producing genetic diversity. We also used the Margalef index to quantify the genetic diversity of the infected group to estimate virus host range. We compared the diversity over the last 250 time steps in intervals of 50-time steps to identify variables affecting host range and under what conditions pathogens evolve as specialists or generalists (Figure 3.13). We observed that pathogens become either specialists or generalists dependent primarily on three parameters: protein binding complexity, survival rate for offspring from infected parents, and the matching threshold. For example, as receptor binding complexity increases, viruses tend to become specialists, which directly relates to the number of targeted receptors due to the difficulty in this case for binding multiple receptors (Figure 3.13a~c). Also a lower survival rate for offspring from infected parents narrows the host range and leads viruses to become specialists because this condition causes the host population size to decay and thus reduces variations within the host population (Figure 3.13d~f). For the same reason, since a low matching threshold is beneficial for virus entry when the binding complexity is low (short L), viruses become specialists (Figure 3.13g~i).

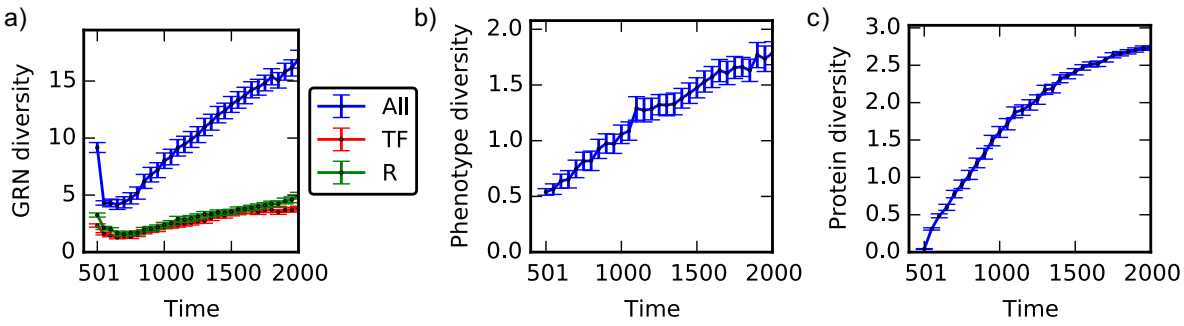


Figure 3. 12. Increased genetic diversity in the gene regulatory networks, phenotypes and receptor proteins. Genetic diversity is measured using the Margalef index (see the last section in Results). a) whole GRNs (blue), transcription factor regulation sub-networks (red), receptor regulation sub-networks (green) of susceptible hosts. b) Phenotypes (gene expression levels) of susceptible populations. c) Receptor sequence of susceptible populations.

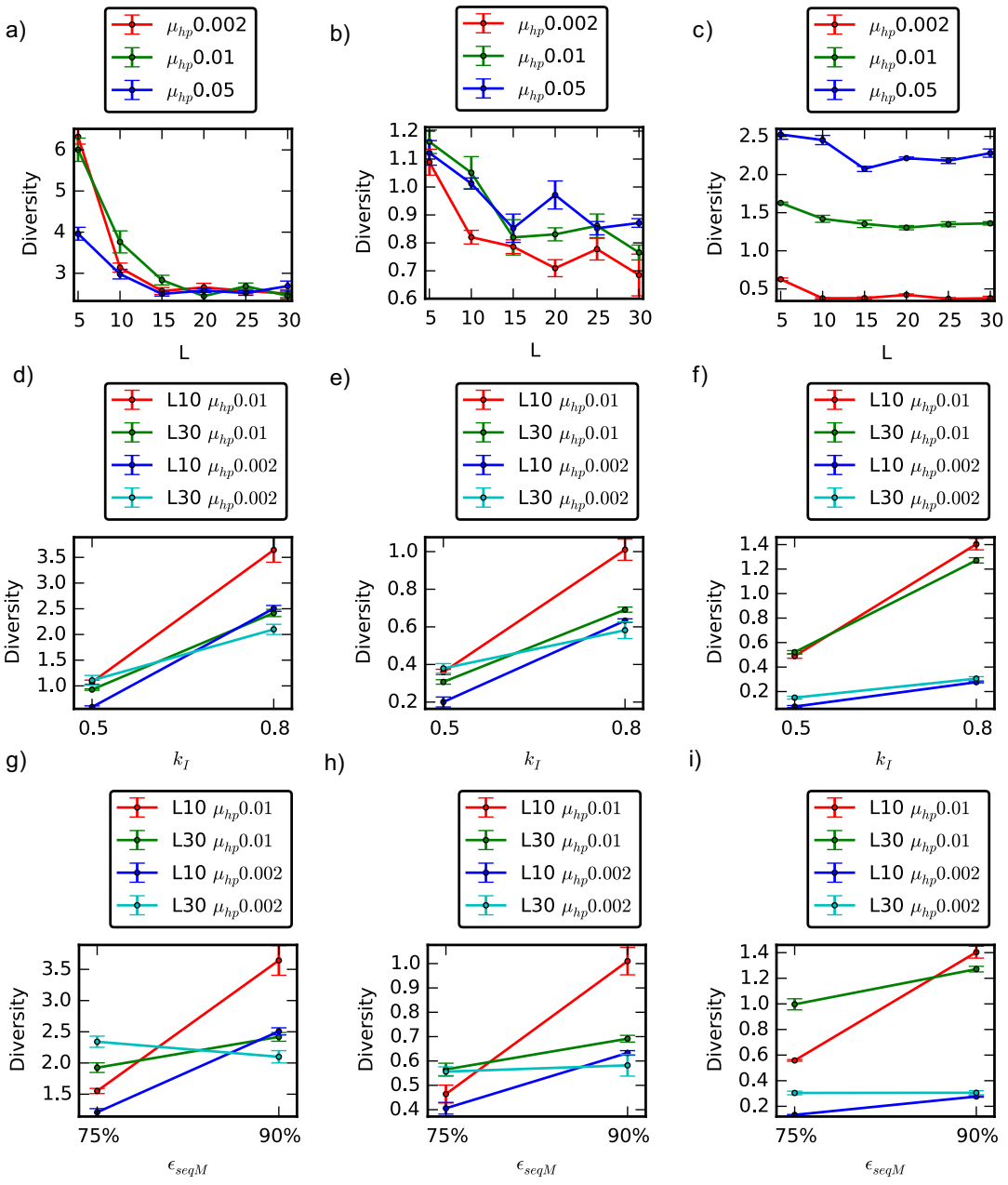


Figure 3. 13. Host range measured by infected host population's genetic diversity under different conditions. The first column is the gene regulatory network diversity, the second column is the phenotype diversity and the last column is the receptor protein sequence diversity. Viruses become specialists when receptor binding complexity ($\boldsymbol{L}$) increases (a,b,c), survival rate for offspring from infected parents ($\boldsymbol{k_I}$) decreases (d,e,f) and amino acid matching threshold for protein binding ($\boldsymbol{\epsilon_{seqM}}$) decreases (g,h,i). For low $\boldsymbol{\epsilon_{seqM}}$ and $\boldsymbol{k_I}$, population dynamics generally follows that shown in Figure 3.2b. Hence, in d~i) we considered all 100 simulations for measuring the genetic diversity.

3.4. Methods

3.4.1 Measure of unevenness among targeted receptors

Every 50 time steps after the coevolution phase has begun, we use the Gini coefficient to calculate unevenness in the targeted receptors among the newly infected hosts. Let y_i $(i = 1, \dots, N_R)$ be the mean number of newly infected hosts who match their sequences to the i^{th} receptor throughout the simulation. If these values are sorted in ascending order such that $y'_1 \leq y'_2 \leq \cdots \leq y'_{n-1} \leq y'_n$, then the Gini coefficient = $\left(n + 1 - 2\frac{\sum_{i=1}^{n} y'_i(n+1-i)}{\sum_{i=1}^{n} y'_i}\right)/n$. Gini coefficient is 1 for the maximum unevenness (inequality) and 0 for perfect evenness (equality).

3.4.2. Measure of ability to switch multiple receptors using gene regulatory network rewiring

Every regulatory interaction in the GRN is mutated 50 times and we measure how often it switches expression of more than one gene. We then measure the average fraction of such perturbations that caused a multi-receptor expression switch over all regulatory interactions in the network for all susceptible individuals.

3.5. Discussion and conclusion

We showed that regulatory changes can be used to suppress expression of cell surface receptor genes leading to a blocking of virus entry. Changes in the expression of virally-targeted receptors has been shown to block virus transmission experimentally, for example, in both

dengue virus (DENV) [100] and Hepatitis C virus (HCV) [99], siRNAs can be used to eliminate cell surface receptors and suppress virus entry and infection. At the same time, specific receptors can be intentionally expressed in the context of tumor gene therapy, for example, allowing adenovirus vectors to be used [106, 107] to deliver apoptosis-activating genes to kill tumor cells.

Two mechanisms of resistance were addressed in our model: rewiring of gene regulatory networks and receptor binding site mutations. The balance in usage between these two mechanisms depends on various conditions. As the protein-protein interaction at the cell surface increases in complexity (in our model represented by the binding site length), viruses tend to target a specific receptor and hosts preferentially use network rewiring more often than receptor amino acid changes. In contrast, when the receptor binding site has lower complexity, viruses are able to enter via multiple receptors and hosts evolve receptor amino acid changes to escape viral protein binding. One can ask why is it that in nature, examples of resistance via receptor amino acid mutations appear to be more common than network rewiring? In the examples of dengue virus (DENV) and hepatitis C virus (HCV) resistance through experimentally-induced receptor down-regulation it was shown that, since there several alternative receptors expressed on the cell surface that viruses can use to enter host cells, multiple inhibitory siRNAs for different receptors worked better than a single siRNA for one receptor, although both studies showed that it was difficult to block infection completely [99]. Thus, for example, HCV can enter human liver cells via several cell surface receptors including CD81 tetraspanin, claudin1(CLDN1), low density lipoprotein receptor receptor (LDLR) and scavenger receptor class B type 1 (SR-B1). In our model, when receptor binding has low complexity, multiple receptors are targeted by viruses and receptor amino acid mutations are used preferentially over network rewiring. Given this observation, the capability of viruses to use alternative receptors for host cell entry is a plausible

explanation of why resistance using network rewiring changes is difficult in practice. Another possible reason for more frequent protein level resistance could be related to the level of functional redundancy among receptors. Higher N_{ER} indicates less functional redundancy among receptors, and we found that protein level resistance was favored for higher N_{ER} (Figure 3.9a). Although functional redundancy is often observed in receptors such as nectin and chemokine receptors as described in Introduction, it is plausible that viruses evolve to target receptors whose absence cannot be compensated for, so that hosts have to express all (or nearly all) required receptors for their normal function, which makes it difficult to use network level resistance.

In order to investigate the importance of including the complex GRN for controlling receptor gene expression, we compared our model with one that did not contain gene regulatory interactions for receptor coding genes. We designed this model by using a diagonal matrix regulatory network both for TF genes and for the receptor coding genes. Complex gene regulation by TFs were removed by having a diagonal matrix with 1s for the regulatory gene network. To satisfy the minimum number of required expressed receptors (N_{ER}/N_R=3/5), we set the initial density of non-zeros on the diagonal for the receptor coding genes with probability 0.7. Here, mutations can occur only on the diagonal of receptor coding genes and no regulation from other genes is possible. Compared to this model, the benefit of having a complex GRN is that the network is capable of evolving increased potential for resistance using network rewiring as shown in Figure 3.11a for complex protein binding (long L), as an example. Here, in the case of complex protein binding where a specific receptor is targeted, it is not possible for the potential for resistance to change because there is only a single entry on the diagonal which can change the expression of the targeted receptor. We compared the preference for GRN level resistance between these two models. We found that the preference of GRN rewiring decreased for the

model without gene regulatory interactions (Figure 3.14a). Furthermore, in order to express at least N_{ER} receptors for the normal host cell function, down-regulating a receptor gene for resistance can be deleterious, and therefore, hosts need to be able to change the expression of multiple receptors simultaneously, in particular to compensate for receptor down-regulation. We found that the systems with complex GRNs evolve the ability to switch the expression of multiple receptors (Figure 3.14b and Methods), whereas without the GRNs, multiple receptor expression change is impossible given a single mutation.

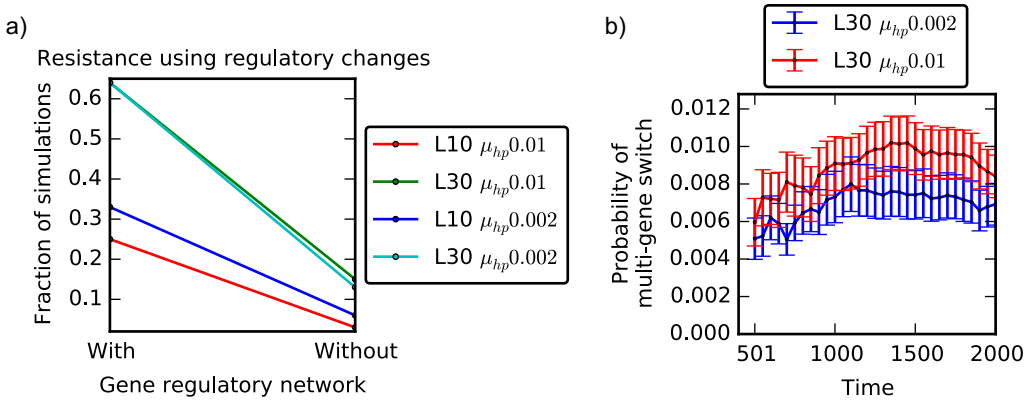


Figure 3. 14. The effect of having a complex gene regulatory network (GRN) for controlling receptor gene expression. a) Preference for resistance using GRN rewiring to protein mutations decreases when there are no regulatory interactions between genes (without regulatory interactions in the gene network) ($\boldsymbol{N_{ER}/N_R}$=3/5, $\boldsymbol{\epsilon_{seqM}}$=90%, $\boldsymbol{k_I}$=0.8). b) The ability to switch the expression of multiple receptors with a complex GRN. The probability of multiple receptor gene expression switching (see Methods) increases during host-virus coevolution ($\boldsymbol{L}$=30, $\boldsymbol{\mu_{hp}}$=0.01 and 0.002, $\boldsymbol{\epsilon_{seqM}}$=90%, $\boldsymbol{k_I}$=0.8).

Although defending from infection at the level of virus entry would appear to be an effective resistance mechanism, the host evolution rate is usually too slow relative to most virus populations and furthermore, viruses are often capable of entering host cells via interaction with multiple receptors. For these reasons, host strategies may have evolved preferentially to allow

viruses to enter cells but to focus defense mechanisms to the post-entry stage by evolving innate and adaptive immune systems. For example, a previous study of North American house finches showed rewiring of gene regulatory networks to up-regulate immune related genes in a relatively short timespan of just 12 years [83].

In addition to network rewiring and receptor amino acid mutations, mutations causing premature stop codons can be used by hosts to block virus entry. CCR5 (CC-chemokine receptor-5) is a co-receptor for HIV entry that facilitates virus entry. A CCR5 allele carrying a 32-bp deletion (ccr5Δ32) in the open reading frame generates a premature stop codon leading to an inactive receptor protein [117, 118]. Homozygous ccr5Δ32/ccr5Δ32 carriers show high immunity to HIV infection and heterozygous wt/ccr5Δ32 carriers show partial resistance to HIV cell entry or delayed progression of the disease. A similar example is an allele of the TVB^R receptor involving a 4-bp insertion which contains a stop codon resulting in protection against Avian Sarcoma and Leukosis Virus (ASLV) entry in chicken [119]. Of note is that even though these stop codon-containing alleles can block virus entry, they work effectively only in homozygous form, in contrast to alleles encoding regulatory repression, which may be effective in single copy form.

Chapter 4. Evolution of environmental robustness in host innate immune systems induced by host-virus interaction

4.1. Background

Infectious disease modeling has been used extensively in the past to understand the dynamics of various pathogens, but rarely do such models address changes at the genetic level. Experimental approaches to study host-virus coevolution are challenging due to the large timescales involved and large population sizes of the co-evolving host and pathogen populations. Instead, computational modeling approaches can be adopted to gain insights into the evolution of infection and resistance mechanisms during host-virus coevolution. In Chapter 3, we studied different resistance strategies at the levels of both GRN and protein interaction in the context of virus entry to the host cell which is the first step of virus infection. Due to the extremely fast virus evolution rate at the protein level, defense at the viral entry level is limited for host populations. Hence, we addressed the hypothesis that fast viral evolution may drive more complex resistance mechanisms such as innate or adaptive immunity. In this chapter, we adopted a computational modeling approach to investigate aspects of the evolution of innate immunity.

Once viruses are detected as they enter host cells, hosts initiate innate immune signaling cascades to reach an antiviral state. Protein Recognition Receptors (PRR), such as Toll-like receptors (TLR) recognize pathogens entering the host cells via detecting Pathogen-associated molecular patterns (PAMP) [120-122]. The PAMP-PRR interaction initiates innate immune signaling cascades in insects and mammals by regulating transcription factors (TFs) such as NF-κB, interferon regulatory TFs such as IRF3 and IRF7 in order to activate pro-inflammatory cytokines which regulate inflammatory responses and induce cells to establish the antiviral state [120, 121].

However, the innate immune pathways are frequently interrupted by various viral proteins (virulence factors). As viruses enter host cells, their virulence factors interact with host components involved in the signaling pathways in order to evade host immunity in various ways [123-125]. For example, in human, suppression of type I interferon (IFN), which has a critical role in inducing many antiviral genes, is an effective immune evasion strategy that has been observed in different viruses including Hepatitis B virus (HBV) [126, 127]. HBV core protein binds to IFN-β (type I IFN) and represses IFN-β expression. IRF3 is another critical TF required for the type I IFN activation. IRF3 is activated by interaction between TBK1/IKKϵ and DDX3 which is interrupted by a HBV polymerase. In addition to these two strategies, the HBV has more modes of Type I IFN suppression [127].

NF-κB is a TF that mediates expression of cytokine genes including type I IFN and a chemokine called Interleukin-8 (IL-8) in macrophages. A chemokine is a cytokine which is able to regulate nearby cells by inducing chemotaxis, a chemical signal that induces cell movement. IκB is an inhibitor protein of the NF-κB. African swine fever virus (ASFV) produces the IκB homologue encoded by A238L gene. Therefore, A238L prevents NF-κB from binding the IL-8 promoter, represses IL-8 expression, and interferes with the inflammatory pathway [128, 129]. A238L is an example of pathogen mimicry of a host regulator protein involved in the immune pathway. Similarly, pathogens mimic host ligands and compete to bind host receptors to evade immunity. For example, Axl is a receptor tyrosine kinase and Gas6 is a Axl binding protein. The Axl/Gas6 pathway has been suggested to be important for IL-15 induced human NK-cell development [130]. *Simian polyoma* virus SV40 protein VP1 mimics Gas6. The VP1-Axl interaction has been suggested to induce the SV40 entry and infection [131]. As another example of host ligand mimicry, NS5 is a Dengue Virus (DENV) protein that mimics a host ligand that

binds to STAT2. As type I IFNs (IFN-α/β) bind Type I IFN receptors on the surface of infected cell, the JAK/STAT signaling pathway is initiated. This signaling pathway activates hundreds of IFN-α/β-stimulated regulatory elements and induces many IFN stimulated genes that are responsible for the antiviral state. Thus, the NS5-STAT2 binding results in STAT2 degradation and blocking of the JAK/STAT signaling pathway [132].

It has been shown that the virus evasion via host factor mimicry induces coevolution between hosts and parasites at the molecular level [78]. The arms race between Poxvirus and human Protein kinase R (PKR) is an example. Here, eIF2α is a critical host protein that initiates protein synthesis. Poxvirus encodes the K3L protein which mimics eIF2α. PKR recognizes double stranded RNA (dsRNA) viruses such as Poxvirus and phosphorylates eIF2α in order to prevent virus protein production. A PKR domain that directly contacts eIF2α changes to compete and defend the K3L binding the PKR [78, 133]. More generally, previous studies found that innate immune-related genes were under positive selection and that innate immune-related proteins evolved rapidly compared to random genes excluding pathogen recognition genes [134-136].

In this chapter, we present a model where we represented a virus as a set of virus proteins (virulence factors) that mimic and bind host regulator proteins such as NF-κB and interferon regulatory factors, which are involved in innate immune pathways. A host individual is represented at two levels: a set of immune-related host proteins that are targeted by virus proteins and a gene regulatory network (GRN) of immune-related genes whose output phenotype determines an antiviral state. In the model, hosts can use two resistance mechanisms: (a) amino acid mutations to defend against virus protein mimicry or binding and (b) tolerating perturbations at the level of the gene regulatory network (GRN) in order to preserve the antiviral

state. We explored how hosts evolve to use these two different levels of resistance. We are interested in understanding how viruses evolve to target host proteins for mimicry or binding, and also how hosts evolve to gain and distribute resistance between the proteins targeted by virus proteins and the GRNs.

4.2. Model

4.2.1. Host innate immunity model and host-virus coevolution model

As viruses enter host cells via cell surface receptors, innate immune signaling pathways are initiated by regulating host regulators which induce inflammatory cytokine production. Consequently, various cytokine-stimulated genes are expressed and they result in an antiviral state in the host cells. Viruses interrupt the innate immune pathways at various levels through protein-protein interactions. A pro-inflammatory signaling pathway is represented as a gene regulatory network (W) of size $N \times N_{TF}$ where N_{TF} is the number of host regulators and N is the total number of genes including the N_{TF} regulators and N_C genes that are responsible for establishing the antiviral state. The individual gene regulatory network (GRN) structure and gene expression dynamics largely follows the original gene regulatory network evolution model (see Chapter 1) [17, 110, 111]. With a given network density (c), each nonzero w_{ij} element in the W matrix is drawn from the Normal distribution, $N(0,1)$. Each row i in W represents the *cis*-regulatory elements of the i^{th} gene.

In the model, there are two different initial regulatory gene expression vectors for a virus-exposed/entered state and for a virus-unexposed state. Once a virus enters a host cell, the host individual sets the initial gene expression level $S(0)$ to $S^I(0)$, a length N_{TF} vector. Following the

gene expression dynamics, $S(t+1) = Sig(W \cdot S^{TF}(t))$, where $Sig(x) = \frac{1}{1+e^{-ax}}$ (a=100), the stable gene expression of all N genes (phenotype) including cytokine-stimulated genes shaping an antiviral state can reach at $\hat{S}^I$, a length N vector. Similarly, at the normal state without viruses, $S(0)$ is set to $S^U(0)$, a length N_{TF} vector and the phenotype is noted as $\hat{S}^U$, a length N vector obtained using the same gene expression dynamics above. In order to differentiate virus -exposed and -unexposed states, we set $S^U(0) \neq S^I(0)$ and $\hat{S}^U \neq \hat{S}^I$. As shown in the Figure 4.1, the top $N_{TF} = 3$ genes are input regulatory genes and the bottom $N_C = 3$ genes are antiviral genes. For a virus-unexposed host individual, a founder individual's phenotype, $\hat{S}^U$ reached from $S^U(0)$ is used for a target phenotype. The target phenotype of the bottom N_C antiviral genes is denoted as $\hat{S}^U_{OPT}$. Similarly, in the virus contact event, the founder individual's phenotype, $\hat{S}^I$ reached from $S^I(0)$ is used for a target phenotype and the target phenotype of the N_C antiviral genes is noted as $\hat{S}^I_{OPT}$.

At the protein level, a host individual possesses N_{TF} regulator proteins in the immune system and a virus possesses N_{VP} number of viral proteins which is not necessarily equal to N_{TF}. Each protein represents a protein binding site as a binary vector of length L, where 0 indicates a polar amino acid and 1 indicates a hydrophobic amino acid. Assuming a pair of two amino acids of the same polarity increases binding affinity, protein-protein binding interaction is assumed to be tight when the percentage of one-to-one amino acid matching among L sites is more than a given threshold (ϵ_{seqM}%). Considering that viruses can disrupt the host immunity via hijacking or mimicking the host regulators, in this model, the host-virus protein-protein interactions represent viral evasion which results in initial state perturbations in the innate immune system, i.e. the perturbed initial regulatory gene expression level from $S^I(0)$ to a different state $S^{I'}(0) \neq$

$S^I(0)$. For N_{TF} host regulators and N_{VP} virus proteins, $N_{TF}\times N_{VP}$ protein-protein interactions occur. For each virus protein, if the virus protein binds to regulator(s), the $S^I(0)$ is perturbed as $s_i^I(0) \to 1 - s_i^I(0)$ at the matched i-th regulator(s). Therefore, the perturbed phenotype of the antiviral genes ($\hat{S}_C^{I\prime}$) is not necessarily maintained closely to the antiviral state, $\hat{S}_{OPT}^I$ due to the initial state disturbance. We measure phenotype distance between $\hat{S}_{OPT}^I$ and perturbed phenotype of antiviral genes, $\hat{S}_C^{I\prime}$ which is reached from the perturbed initial gene expression. Note that the phenotype distance is $\frac{SSD(\hat{S}_{OPT}^I, \hat{S}_C^{I\prime})}{N_C}$, where SSD is the sum of squared distance, and if a phenotype is not reached from the perturbed initial gene expression, we set the phenotype distance 1. Then, we calculate the average phenotype distance across the N_{VP} virus proteins ($d_I = \sum_{i=1}^{N_{VP}} \frac{SSD(\hat{S}_{OPT}^I, \hat{S}_C^{I\prime})}{N_C} / N_{VP}$). Then, if $d_I \geq \epsilon = 10^{-4}$, we assume that antiviral state is not maintained due to disruption by N_{VP} virus proteins, which indicates that the protein level virus evasion causes the host innate immunity to malfunction and the host individual becomes infected.

At the population level, we adopted the SIS model with births and deaths that we used in the previous model in Chapter 3. The susceptible and infected population dynamics are represented using the following difference equations:

$$\Delta S = S(t+1) - S(t) = \eta \cdot b \cdot N(t) \cdot \left(1 - \frac{N(t)}{K}\right) - \xi \cdot \frac{r}{N(t)} \cdot S(t) \cdot I(t) - \lambda_N \cdot S(t) + \gamma \cdot I(t) \quad (1)$$

$$\Delta I = I(t+1) - I(t) = \xi \cdot \frac{r}{N(t)} \cdot S(t) \cdot I(t) - (\lambda_N + \lambda_D + \gamma) \cdot I(t) \quad (2)$$

where $N(t) = S(t) + I(t)$, b=growth rate, K=carrying capacity, $\eta = \frac{\#\ of\ survived\ offspring}{\#\ of\ offspring\ candidates}$, r=contact rate, $\xi = \frac{\#\ of\ infections}{\#\ of\ contacts}$ (determined empirically, as described below), $r \cdot \xi$=transmission rate, λ_N=natural death rate, λ_D=disease related death rate, γ=recovery rate. In the

growth term, $\eta \cdot b \cdot N(t) \cdot \left(1 - \frac{N(t)}{K}\right)$, $b \cdot N(t) \cdot \left(1 - \frac{N(t)}{K}\right)$ is the total number of offspring candidates who function normally under the virus-unexposed condition, i.e., $S^U(0) \rightarrow \hat{S}^U \approx \hat{S}^U_{OPT}$. Among these offspring candidates, only a fraction of the candidates (η) can actually be added to the susceptible population since candidates who have infected parents are less likely to survive. For each host-virus contact event, the virus mutates its amino acid sequences at the point of the infection with mutation rate, $\mu_{vp} = 0.1$ per protein and each virus protein attempts to mimic/bind host proteins. In the infection term, $\xi \cdot \frac{r}{N(t)} \cdot S(t) \cdot I(t)$, among the total number of contacts, $\frac{r}{N(t)} \cdot S(t) \cdot I(t)$, only a fraction ($\xi$) of the contact events lead to actual transmission and host individuals move to infected group if their immune systems (W matrices) do not tolerate the direct perturbation by viruses, i.e., $d_I \geq \epsilon$. On the other hand, if the virus disruption does not lead to a non-antiviral state, i.e., $d_I < \epsilon$, host individuals remain in the susceptible group (Figure 4.1).

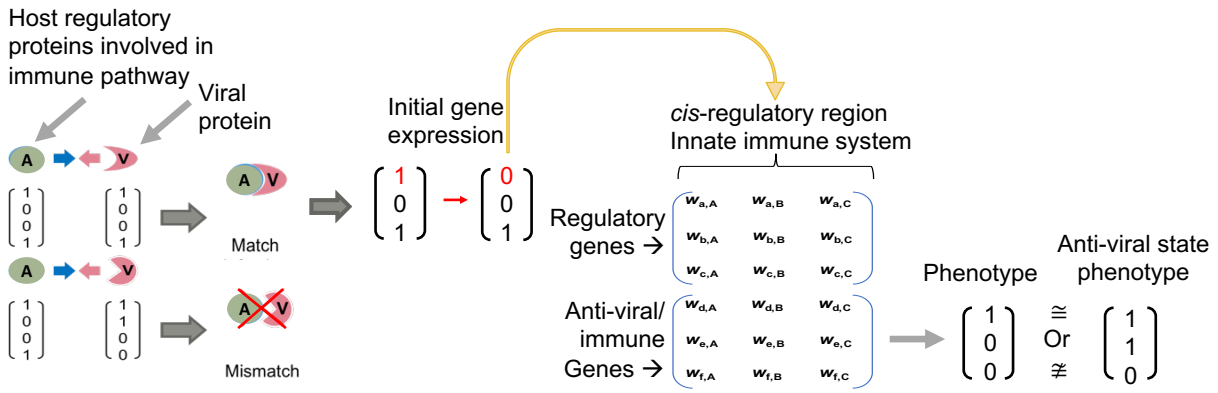


Figure 4. 1. A diagram of host-virus protein-protein interaction and a scheme of host innate immunity interruption by the virus. Host-virus protein-protein interaction is simplified as a one-to-one hydrophobic (0) or polar (1) amino acid matching at a protein binding site. If host regulators are targeted by a virus protein, the initial gene expression of the regulator genes are perturbed. The host individual is infected by the virus if the phenotype from the perturbed initial state is not maintained closely at the target phenotype which represent an anti-viral state. If not, the host individual resists infection and stays in the susceptible group.

4.2.2. Parameters

At the level of the individual, model parameters include the number of host regulator proteins (N_{TF}), the number of virus proteins (N_{VP}), the number of genes that are responsible for establishing the antiviral state (N_C), protein binding site amino acid sequence length (L), host protein mutation rate (μ_{hp}), virus protein mutation rate (μ_{vp}), amino acid matching threshold for receptor binding (ϵ_{seqM}), network density (c), and mutation rate per matrix W (μ) with interaction addition (ρ), deletion (ϕ), and modification (δ) with $\phi + \delta = 1$. Note that $\phi + \delta = 1$, since for an interaction (w_{ij}), deletion and modification are conditional on the interaction being nonzero value ($w_{ij} \neq 0$).

There are also parameters at the level of the population dynamics including offspring survival probability from both infected parents (k_I), disease-related death rate (λ_D), initial host population size M_{initH}, initial virus population size M_{initV}, carrying capacity K, growth rate b, natural death rate λ_N, disease-related death rate λ_D, recovery rate γ, and host-virus contact rate r.

In order to investigate the effect of parameter changes on the evolution of host resistance, we tested a range of parameters above. In Table 4.1, we summarize the range of parameter sets that we used in this study. The default values for the population dynamics related parameters are chosen to make a steady state host population size large enough to investigate evolutionary mechanisms. The individual level parameters for GRN modeling and evolution are chosen based on our previous study [34].

Table 4. 1. The list of model parameters at both the level of population dynamics and at the individual level in symbols with descriptions and parameter values used in this study.

Parameter symbol	Description	Values

L	Protein binding site amino acid sequence length	10, 30
μ_{hp}	Host protein mutation rate	0.003, 0.01
μ_{vp}	Virus protein mutation rate	0.03
N_{TF}	The number of host regulators	6
N_{VP}	The number of virus virulence factors	6
N_C	The number of genes establishing the anti-viral state	6
ϵ_{seqM}	Amino acid matching threshold for receptor binding	90%, 75%
t_{init}	Time at which coevolution begins	1, 601
k_I	Offspring survival probability from both infected parents	0.5, 0.8
ξ	$\frac{\#\ of\ infections}{\#\ of\ contacts}$	Self-determined during simulations
η	$\frac{\#\ of\ survived\ offspring}{\#\ of\ offspring\ candidates}$	Self-determined during simulations
K	Carrying capacity	1000
M_{initH}	Initial host population size	150
M_{initV}	Initial virus population size	5
b	Growth rate	0.15
λ_N	Natural death rate	0.09
λ_D	Disease-related death rate	0.06
γ	Recovery rate	0.05, 0.2
r	Host-virus contact rate	2
c	Network density	0.4
μ	Mutation rate per gene regulatory network	0.1
ϕ	Conditional rate of interaction deletion in gene regulatory network	0. 042
δ	Conditional rate of interaction modification in gene regulatory network	0. 958
σ	Selection pressure	0.1
a	Gene expression mapping sigmoid function parameter	100

4.3. Results

4.3.1. Virus strategy for infection

In the model, a founder virus protein is generated as a copy of a random host protein. For the case of $N_{VP} = N_{TF}$, a founder virus protein of index i is copied from a random host protein of the same index i. Thus, it is expected that each virus protein can bind/mimic a host protein at the beginning of coevolution. A successful virus protein interaction indicates that the virus will perturb the host immune system to potentially disrupt an antiviral state. In order to measure the viral protein ability to target a host protein, we first observed the fraction of infected individuals whose i^{th} protein ($i = 1, \cdots N_{TF}$) is targeted by the j^{th} viral protein ($j = 1, \cdots N_{VP}$). In figure 4.2.a, each cell in a j^{th} subplot indicates $\frac{\#\ of\ infected\ hosts\ whose\ i^{th}\ protein\ is\ targeted\ by\ the\ j^{th}\ viral\ protein}{\#\ of\ infected\ hosts \cdot N_{TF}}$. For simplicity, we call it $V_{i,j}$. The height of stacked cells at a time indicates the average frequency of a host protein being targeted by the j^{th} viral protein. For simplicity, we call it V_j. Then, we averaged the measurements over all viral proteins which indicates the average viral protein ability to target a host protein. For simplicity, we call it $ave(V_j)$. Considering how a founder virus is generated as described above, $ave(V_j)$ value is expected to be close to $\frac{1}{N_{TF}}$ at the beginning. In figure 4.2.c, since we used $N_{TF} = 6$, the $ave(V_j)$ value at the beginning of coevolution is close to $\frac{1}{6}$. Since both host and virus accumulate amino acid mutations, the average viral protein ability to target a host protein will change during coevolution. We found that hosts evolve to evade regulatory protein mimicry during the coevolution, as we observed the $ave(V_j)$ value decreased over time (Figure 4.2b, c). Since the average frequency of a host protein being targeted by a viral protein (V_j's) change unpredictably at early time points, we focused on the end of the simulations. As shown in Figure 4.2d, the ability to mimic/bind a host protein evolves to be very different across

virus proteins, and only a subset of virus proteins evolves to be able to target host proteins, which indicates that specfic virus(es) evolve to target host proteins and others are blocked by host's protein-level defense. An area below a curve in Figure 4.2d relates to overall viral ability to target host proteins. Next, we show that the area changes depending on different model parameters.

We found that the overall viral ability to target host proteins can change depending on the receptor binding complexity (protein sequence length, L), protein binding threshold (ϵ_{seqM}), and recovery rate (γ) (Figure 4.3). As protein binding complexity increases (longer L), only a single viral protein evolves to target host regulator(s) and hosts evolve to escape binding all other viral proteins. An area below a curve is smaller and the overall viral ability to target host proteins decreases for longer L condition (Figure 4.3a vs. Figure 4.3b). A low amino acid matching threshold is beneficial for viruses to target host proteins and induces them to target multiple proteins. Thus, for lower ϵ_{seqM} the area below the curve increases and the overall viral ability to target host proteins increases (Figure 4.3c vs. Figure 4.3d). For a low recovery rate, infected hosts have lower chance to re-enter the susceptible group and remain in the infected group. This induces viruses to accumulate more mutations, which is beneficial to target various host proteins. Hence, low recovery rate increases the overall viral ability to target host proteins (Figure 4.3e vs. Figure 4.3f).

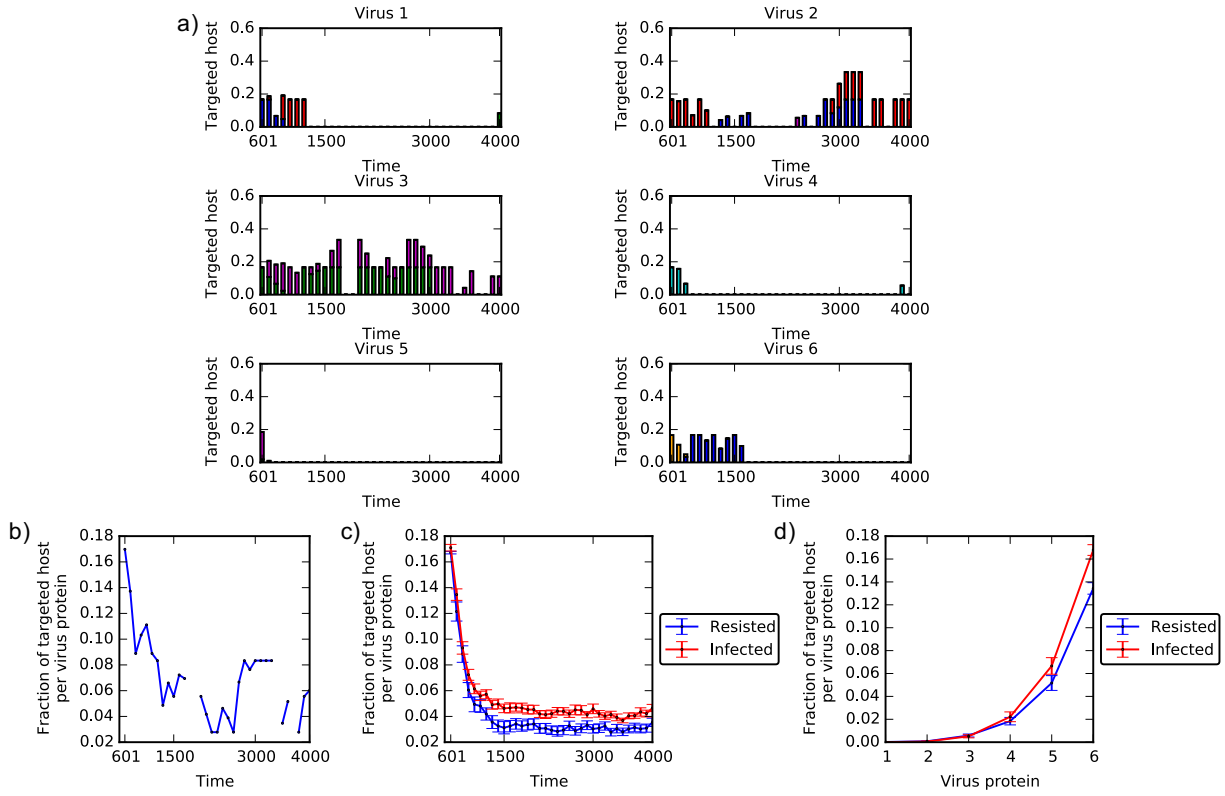


Figure 4. 2. Changes in the distribution of targeted host regulators per virulence factor of a virus. a) The distribution of targeted host proteins within the infected host group per virus protein over time for a single simulation. In each subplot, each colored bar indicates a different host protein. Height of stacked cells ($\boldsymbol{V_j}$ in the main text) at a time point indicates the average frequency of a host protein being targeted by the specific viral protein. The maximun height for each cell is $\frac{\mathbf{1}}{\boldsymbol{N_{TF}}}$ and the range of y-axis is $[\mathbf{0}, \mathbf{1}]$). b) The average viral protein ability to target a host protein. ($\boldsymbol{ave(V_j)}$ in the main text). We averaged $\boldsymbol{V_j}$ values in a) over all $\boldsymbol{N_{VP}}$ virus proteins. c) Red curve: Average of the measurement in b) over 50 simulations. Blue curve: The corresponding measurement for the resisted host group. (error bar: SEM) b) and c) show that hosts use amino acid mutations at protein binding sites to block interactions with virus proteins. d) At the end of a simulation, we sorted virus proteins by $\boldsymbol{V_j}$ values in a) for all 50 simulations and averaged them for resisted (blue) and infected (red) host groups separately. The average of the sorted $\boldsymbol{V_j}$ values at the end of simulations are very different across virus proteins showing that specfic virus(es) evolve to target host proteins and others are blocked by host's protein-level defense.

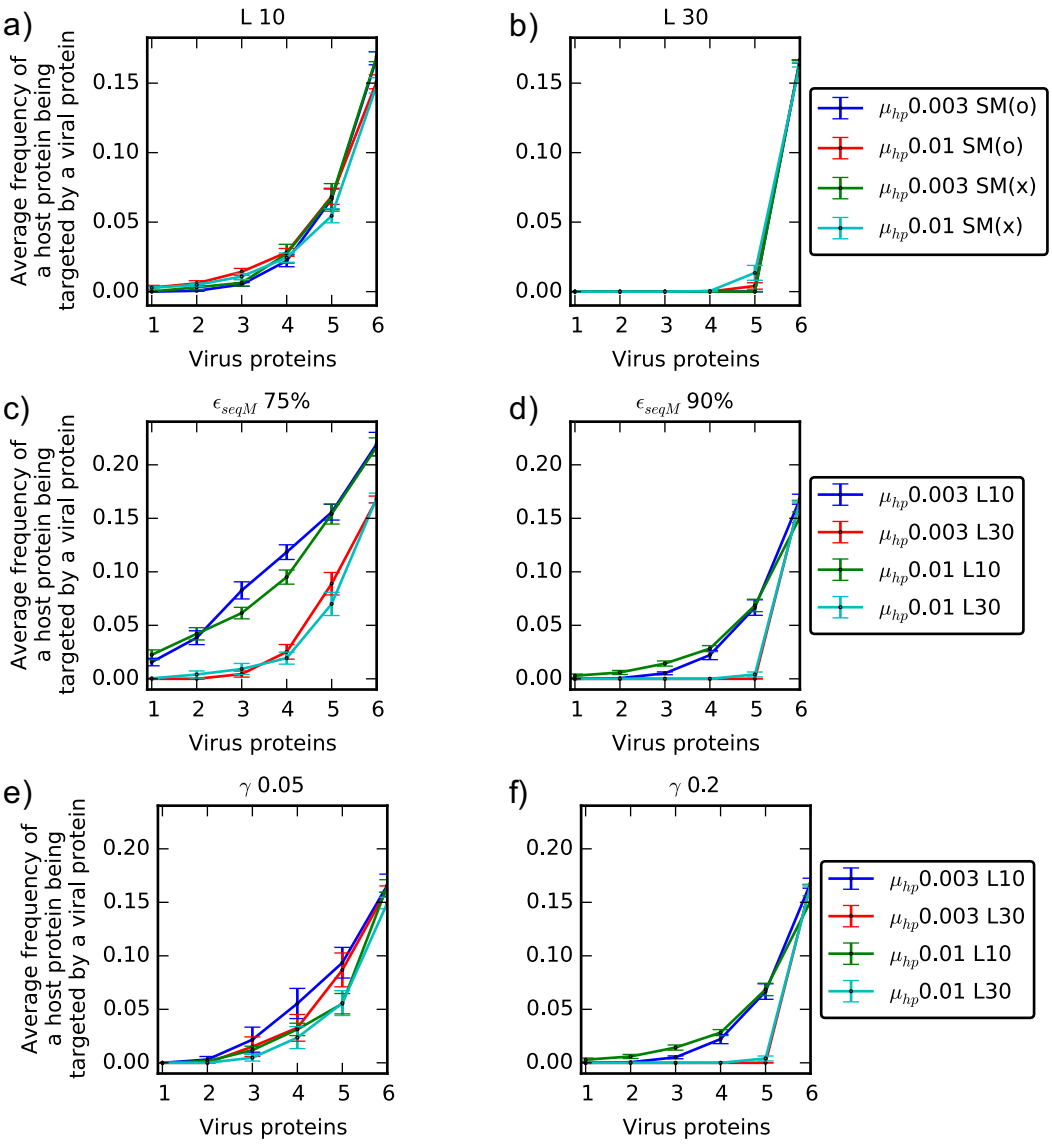


Figure 4. 3. The Distribution of sorted fractions of targeted host proteins among N_{VP} virus proteins for different model parameters. We compared the average of the sorted V_j values at the end of simulations for the infected host group across different a, b) protein binding complexity (L), c, d) protein binding threshold (ϵ_{seqM}), and e, f) recovery rate (γ) (error bar: SEM). For example, in a) and b), the effect of protein binding complexity (L) is investigated for four different parameter sets changing a host protein mutation rate (μ_{hp}) and a state of early stabilizing selection induced by GRN mutations (SM(o/x)). Host proteins are less targeted by viruses and an inequality in the ability to target host proteins among virus proteins is higher for higher protein binding complexity (L) (a vs. b), higher protein binding threshold (ϵ_{seqM}) (c vs. d), and higher recovery rate (γ) (e vs. f).

4.3.2. Two different resistance strategies at the protein interaction level and at the GRN level

Viruses evade host immunity by modulating the innate immune system and disrupting the antiviral state. To defend against viruses, hosts can take two different resistance strategies in the model: 1) phenotypic robustness to environmental perturbations, i.e., initial gene expression changes and 2) amino acid changes in viral protein binding sites to block protein level interactions with viral proteins. Since founder virus proteins are able to bind host regulators, defending the protein mimicry/binding is difficult for the early host population. When the host population is evolved under the stabilizing selection for 600 time points, mutational robustness evolves in GRNs [55]. Previous experimental and theoretical studies suggested the correlation between the mutational robustness and the environmental robustness [5, 10, 137]. Therefore, at the beginning of coevolution, the initially acquired mutational robustness is beneficial for host individuals to use network level resistance strategy to tolerate initial gene expression perturbations by viruses. However, we observed that the early mutational robustness was only beneficial in the early coevolution stage but did not influence on the host resistance and the virus pathogenicity afterwards. Regardless of mutational robustness acquired beforehand, as host individuals accumulate amino acid mutations to avoid protein interaction with viral proteins, they become able to balance both network level and protein level resistance strategies. Here we consider how hosts balance the usage of these two strategies and what conditions determine their relative preference. At each time point, we measure the fraction of individuals that resisted using GRN level resistance by buffering initial regulatory gene expression perturbations rather than protein level resistance which works by avoiding all host protein interactions with virus proteins. We proceed by counting the fraction of exposed hosts to viruses that allow protein mimicry

(protein-protein interaction) and thus direct network level perturbations, but which maintain the antiviral state at every time point. We then calculated the average of these measurements over all time points throughout the simulation to quantify the relative preference for network level resistance over protein level resistance. We specifically measure this at the end of simulation where the fraction of resisted individuals using network level resistance reaches a stable state. We compare these measurements across different parameter sets to figure out under what conditions, network level resistance is preferentially used. We found that the relative preference for network level resistance depends on the receptor binding complexity (protein sequence length, L), host protein mutation rate (μ_{hp}), protein binding threshold (ϵ_{seqM}), and recovery rate (γ) (Figure 4.4). In the Figure 4.4a, we compared the effect of protein binding complexity for different μ_{hp} values and for different early states with (SM(o)) and without (SM(o)) the initial stabilizing selection induced by GRN mutations. Again, as shown in Figure 4.3a and b, an area below a curve in Figure 4.3b (L=10) is bigger than in Figure 4.3a (L=30) indicating that the overall viral ability to target host proteins is higher for shorter L. From the host's point of view, for less complex protein binding (shorter L), more proteins are targeted by viruses, and thus the protein level resistance strategy is less favored. Hence, when protein binding complexity is low, the network level resistance is used more often (Figure 4.4a). When host protein mutation rate (μ_{hp}) is low, it is difficult to catch up with the fast-evolving virus proteins, leading hosts to evolve GRN level resistance instead of using amino acid mutations to evade protein mimicry (Figure 4.4b). Similarly, since a lower amino acid matching threshold (lower ϵ_{seqM}) increases the overall viral ability to target host proteins (Figure 4.3c, d), the GRN level resistance strategy is relatively more favored than the protein level strategy (Figure 4.4.c). In Figure 4.3e and f, it has been shown that low recovery rate (γ) increases the overall viral ability to target host proteins.

The lower γ, therefore, leads to increase the relative usage of the GRN level resistance strategy (Figure 4.4e).

Lastly, it turns out that there is a correlation between the relative preference gene regulatory network (GRN) level resistance strategy shown in Figure 4.4 and the overall viral ability to target host proteins shown in Figure 4.3. For conditions that are beneficial for viruses to target host proteins increasing $ave(V_j)$, hosts allow mimicry by viruses but instead preferentially choose network level resistance by buffering the initial regulatory gene disruption.

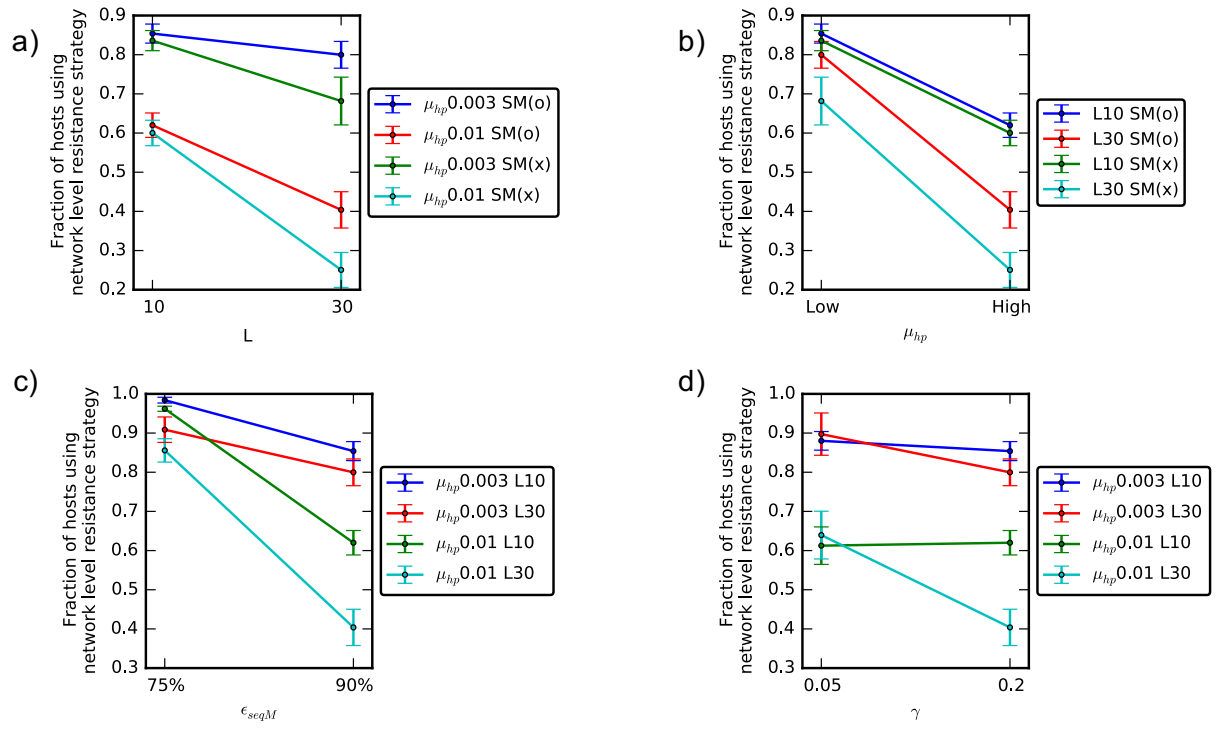


Figure 4. 4. Preference for gene regulatory network (GRN) level resistance strategy rather than amino acid mutations at viral protein binding sites. The GRN level resistance strategy is preferred, and the fraction of resisted individuals that used the GRN level resistance strategy (buffers initial regulatory gene level perturbations by viruses) increases for a) lower protein binding complexity ($\boldsymbol{L}$), b) lower host protein mutation rate ($\boldsymbol{\mu_{hp}}$), c) lower protein binding threshold ($\boldsymbol{\epsilon_{seqM}}$), and d) lower recovery rate ($\boldsymbol{\gamma}$).

4.3.3. Hosts evolve environmental robustness

We observed that host-virus coevolution drives the evolution of environmental robustness, as shown by the increased tolerance to random initial gene expression perturbations (Figure 4.5). In order to measure the environmental robustness, we perturb $S(0)$, each gene at a time and measure the average phenotype distances between the perturbed phenotypes and $\hat{S}_{OPT}^{I}$. Again, if stable gene expression (phenotype) is not reached, we set the phenotype distance to 1. In the previous section, we found that the overall ability of viruses to target host proteins was positively correlated with the relative preference for the GRN level resistance strategy. As more host proteins are targeted by viruses, the initial regulatory gene state becomes more different from the normal state. This makes the host population come under stronger selection for environmental robustness. Therefore, under these conditions that lead to a higher fraction of targeted host proteins per virus protein, hosts both use GRN level resistance more frequently and evolve higher environmental robustness (Figure 4.6).

In the previous section (Figure 4.3), we specified model parameters that change the viral ability to mimic host proteins. The more host proteins being targeted lead to more GRNs being perturbed by initial regulatory gene changes. Consequently, those conditions for the host population to suffer more network perturbation facilitate them to evolve higher level of environmental robustness (Figure 4.6).

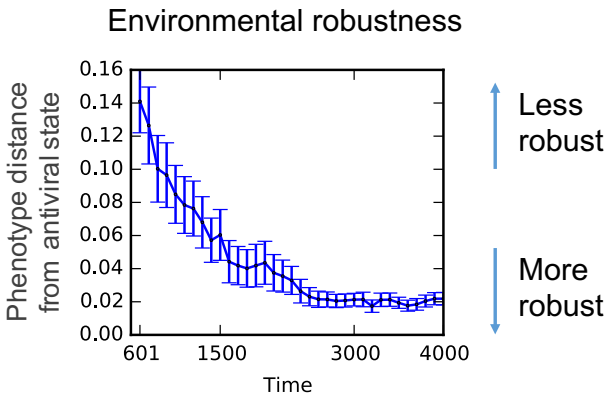


Figure 4. 5. Environmental robustness increases during host-virus coevolution. We randomly perturb the initial regulatory gene expression and measured phenotype distance between $\widehat{\boldsymbol{S}}_{\boldsymbol{OPT}}^{\boldsymbol{I}}$ and perturbed phenotype of antiviral genes, $\widehat{\boldsymbol{S}}_{\boldsymbol{C}}^{\boldsymbol{I}'}$. The phenotype distance imposed by environmental fluctuation decreases over time which indicates increased environmental robustness. ($\boldsymbol{L}$=10, $\boldsymbol{\mu_{hp}}$ =0.003, $\boldsymbol{\epsilon_{seqM}}$=90%, $\boldsymbol{k_I}$=0.8.)

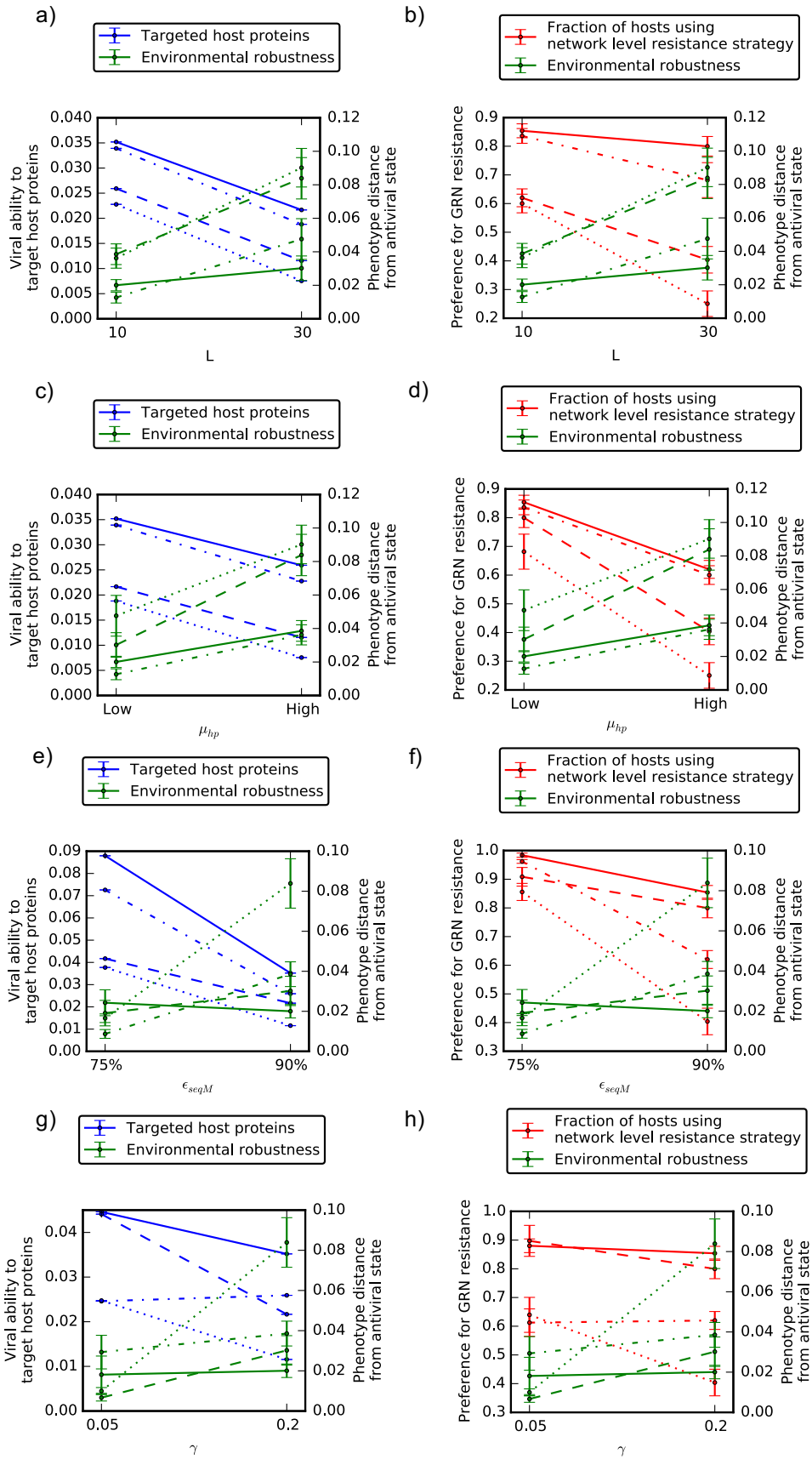

a)
Targeted host proteins
Environmental robustness
Viral ability to target host proteins
Phenotype distance from antiviral state
L
b)
Fraction of hosts using network level resistance strategy
Environmental robustness
Preference for GRN resistance
Phenotype distance from antiviral state
L
c)
Targeted host proteins
Environmental robustness
Viral ability to target host proteins
Phenotype distance from antiviral state
Low
High
μ_{hp}
d)
Fraction of hosts using network level resistance strategy
Environmental robustness
Preference for GRN resistance
Phenotype distance from antiviral state
Low
High
μ_{hp}
e)
Targeted host proteins
Environmental robustness
Viral ability to target host proteins
Phenotype distance from antiviral state
75%
90%
ϵ_{seqM}
f)
Fraction of hosts using network level resistance strategy
Environmental robustness
Preference for GRN resistance
Phenotype distance from antiviral state
75%
90%
ϵ_{seqM}
g)
Targeted host proteins
Environmental robustness
Viral ability to target host proteins
Phenotype distance from antiviral state
γ
h)
Fraction of hosts using network level resistance strategy
Environmental robustness
Preference for GRN resistance
Phenotype distance from antiviral state
γ

Figure 4. 6. A correlation between the overall viral ability to target host proteins and the environmental robustness for initial gene expression perturbations, and a correlation between the preference for the GRN level resistance strategy and the environmental robustness for initial gene expression perturbations. (a, c, e, g) For protein binding complexity ($\boldsymbol{L}$), host protein mutation rate ($\boldsymbol{\mu_{hp}}$), protein binding threshold ($\boldsymbol{\epsilon_{seqM}}$), and recovery rate ($\boldsymbol{\gamma}$), as the viral ability to target host proteins increases, hosts tend to evolve higher environmental robustness for random initial regulatory gene perturbations. (b, d, f, h) Also, for these model parameters, as hosts evolve higher environmental robustness for random initial regulatory gene perturbations, they tend to prefer the gene regulatory network level resistance strategy.

4.4. Conclusion and future work

A previous study has shown that viruses impose coevolutionary selection on their hosts (mammals) and viruses have been suggested to be a critical component that drives protein adaptation of conserved proteins [138]. This work also found that virus-interacting proteins included not only antiviral and immune system proteins but also housekeeping proteins and other proteins with various different functions unrelated to immunity. In our study, we proposed that complexity in innate immune signaling pathways is required to defend against interference by various virus proteins including mimicry of host regulators that have critical functions in signaling pathways. Since viruses evolve extremely fast compared to their hosts, protein level defense against protein mimicry by mutations in virus protein binding sites may not be sufficient. Therefore, hosts may have evolved higher level defense mechanisms in the innate immune pathways. In this study, we found that sabotage by virulence factors of viruses induce selection pressure for environmental robustness in innate immune systems. Furthermore, we observed that the level of the environmental robustness was determined by different conditions that were also related to virus pathogenicity. In nature, hosts have evolved complex immune signaling pathways. In the model, increasing the number of regulators (N_{TF}) involved in the host system

gives viruses more chances to interrupt immunity. Since it also increases size and complexity of the immune systems, we can understand the evolution of complexity in the innate immunity using the model.

The viral evasion of host innate immune systems via protein-protein interactions has been reported in many previous studies [124, 125]. As represented in the model, viral proteins have protein-protein interaction with host TFs to interrupt antiviral gene transcription [125, 139]. However, in some studies of Hepatitis B virus (HBV) and Kaposi's sarcoma-associated herpesvirus (KSHV), it has been shown that viral proteins target not only host proteins but also host gene promoters of various genes involved in signaling pathways [140, 141]. Considering that the viral evasion via viral proteins targeting host gene promoter has not widely been reported for many viruses, only some of viruses may have evolved to use their proteins to bind regulatory regions and change host gene expression. For future work, we intend to investigate the evolution of viral genotypes and proteins to study what factors (e.g. viral genome size) may affect the ability of viruses to adopt those two layers of viral evasion strategies, and to study how they influence on the evolution of pathogenicity and host immune systems.

Viruses target host proteins in order to interfere with innate immune signaling pathways. The host-virus protein-protein interactions inhibit the critical host protein functions involved in the signaling cascades. In this study, we focused on protein-protein interactions of host regulatory proteins and virus proteins. Viruses have reported to target both inhibitory regulatory proteins such as IκB, SOCS (suppressor of cytokine signaling) proteins, and TFAF1 and activating regulatory proteins such as STAT2 and DEAD-box RNA helicase (DDX3) to disrupt innate immunity in TNF signaling pathways and JAK/STAT pathways [142]. As another direction for future work, we intend to investigate the effect of activating vs. inhibitory regulator

targeting on virus pathogenicity and host resistance. Viruses may have evolved to target either activating or inhibitory regulator proteins for higher transmissibility. Likewise, hosts may have a preference for a specific mode of regulator for evolving higher environmental robustness.

Chapter 5. Summary and future work

5.1. Summary and future work

In the three studies above, we found that the hosts evolve to increase complexity in their biological systems depending on the interacting and coevolving parasites (or viruses). In particular, host individuals evolved their gene regulatory networks under the selection pressure imposed by the parasite (or virus) individuals and vice versa.

In Chapter 2, we explored evolutionary features in gene regulatory networks of two different organisms of similar levels of complexity under antagonistic coevolutionary selection pressure. By using two interacting species, key model parameters that determine the fitness landscapes became emergent properties of the model, avoiding the need to impose these parameters externally. Both host and parasite populations could achieve frequent and accurate phenotype changes due to broadly distributed sensitivity in a gene regulatory network. We found that sensitivity follows a pattern, similar to that of the game "whack-a-mole", in which sensitive sites mutate, thus becoming insensitive, but new sensitive sites emerge to take their place. We predict that this type of sensitivity will evolve under conditions of strong directional selection, an observation that helps interpret existing experimental evidence, for example, during the emergence of bacterial antibiotic resistance.

In Chapter 3, we studied the evolution of the first step during a viral infection, where we considered two possible levels of resistance mechanism in the host: 1) at the level of the protein binding interaction between host receptors and a virus protein, and 2) at the level of receptor protein expression regulation where we use a standard gene regulatory network model using our host-virus coevolution model. We explored a range of different conditions (model parameters)

that affect host evolutionary dynamics and, in particular, the balance between the use of different resistance mechanisms. We found that host resistance and viral pathogenicity depend on quite different evolutionary conditions. Viruses may evolve cell entry strategies that use small receptor binding regions, represented by low complexity binding in our model. Our modeling results suggest that if the virus adopts a strategy based on binding to low complexity sites on the host receptor, the host will select a defense strategy at the protein (receptor) level, rather than at the level of the regulatory network - a virus-host strategy that appears to have been selected most often in nature.

After viruses enter host cells, they interfere with innate immune systems via protein-protein interactions such as molecular mimicry of various host proteins involved in the immunity. In Chapter 4, we developed a model of the host innate immunity evolution in the context of host-virus coevolution. In this chapter, we discussed viral mechanisms for pathogenicity and how hosts evolve their defense mechanisms depending on different viral mechanisms. We found that the host evolved to optimize the use of 1) mutations at protein-protein interaction sites to avoid mimicry/binding and 2) environmental robustness in the innate immune systems imposed by viral disruption of the immune systems. For future work, we aim to focus more on the evolution of complexity in host immune systems and viral systems. Using this model, we can study how the complex immune systems evolve to influence viral evasion and host defense strategies. Also, focusing on viruses such as HBV and KSHV, which bind host gene promoters to interrupt host signaling pathways for infection, we intend to study how different levels of viral genome complexity relate to viral pathogenicity and host defense mechanisms.

Bibliography

1. Waddington, C.H., ***Genetic Assimilation of an Acquired Character***. Evolution, 1953. **7**(2): p. 118-126.

2. Waddington, C.H., *Genetic Assimilation of the Bithorax Phenotype.* Evolution, 1956. **10**(1): p. 1-13.

3. Waddington, C.H., *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. 1957, London : Allen & Unwin.

4. Waddington, C.H., *Canalization of Development and Genetic Assimilation of Acquired Characters.* Nature, 1959. **183**(4676): p. 1654-1655.

5. Masel, J. and M.L. Siegal, *Robustness: mechanisms and consequences.* Trends Genet, 2009. **25**(9): p. 395-403.

6. McGuigan, K. and C.M. Sgro, *Evolutionary consequences of cryptic genetic variation.* Trends Ecol Evol, 2009. **24**(6): p. 305-11.

7. Edelman, G.M. and J.A. Gally, *Degeneracy and complexity in biological systems.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13763-8.

8. Proulx, S.R. and P.C. Phillips, *The opportunity for canalization and the evolution of genetic networks.* Am Nat, 2005. **165**(2): p. 147-62.

9. Little, J.W., D.P. Shepley, and D.W. Wert, *Robustness of a gene regulatory circuit.* EMBO J, 1999. **18**(15): p. 4299-307.

10. Szollosi, G.J. and I. Derenyi, *Congruent evolution of genetic and environmental robustness in micro-RNA.* Mol Biol Evol, 2009. **26**(4): p. 867-74.

11. Tokuriki, N. and D.S. Tawfik, *Stability effects of mutations and protein evolvability.* Curr Opin Struct Biol, 2009. **19**(5): p. 596-604.

12. Soyer, O.S. and S. Bonhoeffer, *Evolution of complexity in signaling pathways.* Proc Natl Acad Sci U S A, 2006. **103**(44): p. 16337-42.

13. Barkai, N. and S. Leibler, *Robustness in simple biochemical networks.* Nature, 1997. **387**(6636): p. 913-7.

14. Staton, A.A., H. Knaut, and A.J. Giraldez, *miRNA regulation of Sdf1 chemokine signaling provides genetic robustness to germ cell migration.* Nat Genet, 2011. **43**(3): p. 204-11.

15. Larhlimi, A., et al., *Robustness of metabolic networks: a review of existing definitions.* Biosystems, 2011. **106**(1): p. 1-8.

16. MacCarthy, T., R. Seymour, and A. Pomiankowski, *The evolutionary potential of the Drosophila sex determination gene network.* J Theor Biol, 2003. **225**(4): p. 461-8.

17. Ciliberti, S., O.C. Martin, and A. Wagner, *Robustness can evolve gradually in complex regulatory gene networks with varying topology.* PLoS Comput Biol, 2007. **3**(2): p. e15.

18. Bilgin, T., I.A. Kurnaz, and A. Wagner, *Selection shapes the robustness of ligand-binding amino acids.* J Mol Evol, 2013. **76**(5): p. 343-9.

19. van Nimwegen, E., J.P. Crutchfield, and M. Huynen, *Neutral evolution of mutational robustness.* Proc Natl Acad Sci U S A, 1999. **96**(17): p. 9716-20.

20. Isalan, M., et al., *Evolvability and hierarchy in rewired bacterial gene networks.* Nature, 2008. **452**(7189): p. 840-5.

21. Edwards, J.S. and B.O. Palsson, *Robustness analysis of the Escherichia coli metabolic network.* Biotechnol Prog, 2000. **16**(6): p. 927-39.

22. Smart, A.G., L.A. Amaral, and J.M. Ottino, *Cascading failure and robustness in metabolic networks.* Proc Natl Acad Sci U S A, 2008. **105**(36): p. 13223-8.

23. Behre, J., et al., *Structural robustness of metabolic networks with respect to multiple knockouts.* J Theor Biol, 2008. **252**(3): p. 433-41.

24. Rutherford, S.L. and S. Lindquist, *Hsp90 as a capacitor for morphological evolution.* Nature, 1998. **396**(6709): p. 336-42.

25. Milton, C.C., et al., *Quantitative trait symmetry independent of Hsp90 buffering: distinct modes of genetic canalization and developmental stability.* Proc Natl Acad Sci U S A, 2003. **100**(23): p. 13396-401.

26. Queitsch, C., T.A. Sangster, and S. Lindquist, *Hsp90 as a capacitor of phenotypic variation.* Nature, 2002. **417**(6889): p. 618-24.

27. Nuismer, S.L., P. Jordano, and J. Bascompte, *Coevolution and the architecture of mutualistic networks.* Evolution, 2013. **67**(2): p. 338-54.

28. Guimaraes, P.R., Jr., P. Jordano, and J.N. Thompson, *Evolution and coevolution in mutualistic networks.* Ecol Lett, 2011. **14**(9): p. 877-85.

29. Strauss, S.Y. and R.E. Irwin, *Ecological and Evolutionary Consequences of Multispecies Plant-Animal Interactions.* Annual Review of Ecology, Evolution, and Systematics, 2004. **35**(1): p. 435-466.

30. Wagner, A., *Does Evolutionary Plasticity Evolve?* Evolution, 1996. **50**(3): p. 1008.

31. Bergman, A. and M.L. Siegal, *Evolutionary capacitance as a general feature of complex gene networks.* Nature, 2003. **424**(6948): p. 549-52.

32. Pujato, M., et al., *The underlying molecular and network level mechanisms in the evolution of robustness in gene regulatory networks.* PLoS Comput Biol, 2013. **9**(1): p. e1002865.

33. van Dijk, A.D., S. van Mourik, and R.C. van Ham, *Mutational robustness of gene regulatory networks.* PLoS One, 2012. **7**(1): p. e30591.

34. Shin, J. and T. MacCarthy, *Antagonistic Coevolution Drives Whack-alpha-Mole Sensitivity in Gene Regulatory Networks.* PLoS Computational Biology, 2015. **11**(10).

35. Kilner, R.M. and N.E. Langmore, *Cuckoos versus hosts in insects and birds: adaptations, counter-adaptations and outcomes.* Biol Rev Camb Philos Soc, 2011. **86**(4): p. 836-52.

36. Spottiswoode, C.N. and M. Stevens, *Host-parasite arms races and rapid changes in bird egg appearance.* Am Nat, 2012. **179**(5): p. 633-48.

37. Nash, D.R., et al., *A mosaic of chemical coevolution in a large blue butterfly.* Science, 2008. **319**(5859): p. 88-90.

38. Wagner, A., *Robustness and evolvability: a paradox resolved.* Proc Biol Sci, 2008. **275**(1630): p. 91-100.

39. Draghi, J. and G.P. Wagner, *The evolutionary dynamics of evolvability in a gene network model.* J Evol Biol, 2009. **22**(3): p. 599-611.

40. Whitacre, J.M., *Degeneracy: a link between evolvability, robustness and complexity in biological systems.* Theor Biol Med Model, 2010. **7**: p. 6.

41. Papp, B., R.A. Notebaart, and C. Pal, *Systems-biology approaches for predicting genomic evolution.* Nat Rev Genet, 2011. **12**(9): p. 591-602.

42. Kashtan, N. and U. Alon, *Spontaneous evolution of modularity and network motifs.* Proc Natl Acad Sci U S A, 2005. **102**(39): p. 13773-8.

43. Espinosa-Soto, C. and A. Wagner, *Specialization can drive the evolution of modularity.* PLoS Comput Biol, 2010. **6**(3): p. e1000719.

44. Parter, M., N. Kashtan, and U. Alon, *Facilitated variation: how evolution learns from past environments to generalize to new environments.* PLoS Comput Biol, 2008. **4**(11): p. e1000206.

45. Wagner, G.P., M. Pavlicev, and J.M. Cheverud, *The road to modularity.* Nat Rev Genet, 2007. **8**(12): p. 921-31.

46. Clune, J., J.B. Mouret, and H. Lipson, *The evolutionary origins of modularity.* Proc Biol Sci, 2013. **280**(1755): p. 20122863.

47. Frankel, N., S. Wang, and D.L. Stern, *Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution.* Proc Natl Acad Sci U S A, 2012. **109**(51): p. 20975-9.

48. Cresko, W.A., et al., *Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6050-5.

49. Consortium, H.G., *Butterfly genome reveals promiscuous exchange of mimicry adaptations among species.* Nature, 2012. **487**(7405): p. 94-8.

50. Suerbaum, S. and C. Josenhans, *Helicobacter pylori evolution and phenotypic diversification in a changing host.* Nat Rev Microbiol, 2007. **5**(6): p. 441-52.

51. Wong, A., N. Rodrigue, and R. Kassen, *Genomics of adaptation during experimental evolution of the opportunistic pathogen Pseudomonas aeruginosa.* PLoS Genet, 2012. **8**(9): p. e1002928.

52. Woods, R., et al., *Tests of parallel molecular evolution in a long-term experiment with Escherichia coli.* Proc Natl Acad Sci U S A, 2006. **103**(24): p. 9107-12.

53. Azevedo, R.B., et al., *Sexual reproduction selects for robustness and negative epistasis in artificial gene networks.* Nature, 2006. **440**(7080): p. 87-90.

54. Leclerc, R.D., *Survival of the sparsest: robust gene networks are parsimonious.* Mol Syst Biol, 2008. **4**: p. 213.

55. Siegal, M.L. and A. Bergman, *Waddington's canalization revisited: developmental stability and evolution.* Proc Natl Acad Sci U S A, 2002. **99**(16): p. 10528-32.

56. MacCarthy, T., R.M. Seymour, and A. Pomiankowski, *Differential regulation drives plasticity in sex determination gene networks.* BMC Evol Biol, 2010. **10**: p. 388.

57. Ciliberti, S., O.C. Martin, and A. Wagner, *Innovation and robustness in complex regulatory gene networks.* Proc Natl Acad Sci U S A, 2007. **104**(34): p. 13591-6.

58. Kaneko, K., *Proportionality between variances in gene expression induced by noise and mutation: consequence of evolutionary robustness.* BMC Evol Biol, 2011. **11**: p. 27.

59. Brandes, U., et al., *On Modularity Clustering.* IEEE Transactions on Knowledge and Data Engineering, 2008. **20**(2): p. 172-188.

60. Huerta-Sanchez, E. and R. Durrett, *Wagner's canalization model.* Theor Popul Biol, 2007. **71**(2): p. 121-30.

61. Otto, S.P. and S.L. Nuismer, *Species interactions and the evolution of sex.* Science, 2004. **304**(5673): p. 1018-20.

62. Peters, A.D. and C.M. Lively, *Short- and long-term benefits and detriments to recombination under antagonistic coevolution.* J Evol Biol, 2007. **20**(3): p. 1206-17.

63. Martin, O.C. and A. Wagner, *Effects of recombination on complex regulatory circuits.* Genetics, 2009. **183**(2): p. 673-84, 1SI-8SI.

64. Pineda-Krch, M., *Persistence and Loss of Meiotic Recombination Hotspots.* Genetics, 2005. **169**(4): p. 2319-2333.

65. Jeffreys, A.J. and R. Neumann, *Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot.* Nature Genetics, 2002. **31**(3): p. 267-271.

66. Winckler, W., et al., *Comparison of fine-scale recombination rates in humans and chimpanzees.* Science, 2005. **308**(5718): p. 107-11.

67. Coop, G., et al., *High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans.* Science, 2008. **319**(5868): p. 1395-8.

68. Davies, J. and D. Davies, *Origins and evolution of antibiotic resistance.* Microbiol Mol Biol Rev, 2010. **74**(3): p. 417-33.

69. Gniadkowski, M., *Evolution of extended-spectrum beta-lactamases by mutation.* Clin Microbiol Infect, 2008. **14 Suppl 1**: p. 11-32.

70. Stern, D.L., *Evolution, development, & the predictable genome.* 2011, Greenwood Village, Colo.: Roberts and Co. Publishers.

71. Mckenzie, J.A., *Selection at the Dieldrin Resistance Locus in Overwintering Populations of Lucilia-Cuprina (Wiedemann).* Australian Journal of Zoology, 1990. **38**(5): p. 493-501.

72. McKenzie, J.A. and G.M. Clarke, *Diazinon resistance, fluctuating asymmetry and fitness in the Australian sheep blowfly, lucilia cuprina.* Genetics, 1988. **120**(1): p. 213-20.

73. Ortlund, E.A., et al., *Crystal structure of an ancient protein: evolution by conformational epistasis.* Science, 2007. **317**(5844): p. 1544-8.

74. Phillips, P.C., *Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems.* Nat Rev Genet, 2008. **9**(11): p. 855-67.

75. Parter, M., N. Kashtan, and U. Alon, *Environmental variability and modularity of bacterial metabolic networks.* BMC Evol Biol, 2007. **7**: p. 169.

76. Shin, J. and T. MacCarthy, *Potential for evolution of complex defense strategies in a multi-scale model of virus-host coevolution.* BMC Evol Biol, 2016. **16**(1): p. 233.

77. Sironi, M., et al., *Evolutionary insights into host-pathogen interactions from mammalian sequence data.* Nature Reviews Genetics, 2015. **16**(4): p. 224-236.

78. Daugherty, M.D. and H.S. Malik, *Rules of Engagement: Molecular Insights from Host-Virus Arms Races*, in *Annual Review of Genetics, Vol 46*, B.L. Bassler, Editor. 2012. p. 677-700.

79. Wichman, H.A., et al., *Different trajectories of parallel evolution during viral adaptation.* Science, 1999. **285**(5426): p. 422-424.

80. Woolhouse, M.E.J., et al., *Biological and biomedical implications of the co-evolution of pathogens and their hosts.* Nature Genetics, 2002. **32**(4): p. 569-577.

81. Thrall, P.H., et al., *Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation.* Ecology Letters, 2012. **15**(5): p. 425-435.

82. Barribeau, S.M., et al., *Gene expression differences underlying genotype-by-genotype specificity in a host-parasite system.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(9): p. 3496-3501.

83. Bonneaud, C., et al., *Rapid evolution of disease resistance is accompanied by functional changes in gene expression in a wild bird.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(19): p. 7866-7871.

84. Martiny, J.B.H., et al., *Antagonistic Coevolution of Marine Planktonic Viruses and Their Hosts*, in *Annual Review of Marine Science, Vol 6*, C.A. Carlson and S.J. Giovannoni, Editors. 2014. p. 393-414.

85. van Nimwegen, E., *Influenza escapes immunity along neutral networks.* Science, 2006. **314**(5807): p. 1884-1886.

86. Worobey, M., A. Bjork, and J.O. Wertheim, *Point, Counterpoint: The Evolution of Pathogenic Viruses and their Human Hosts.* Annual Review of Ecology, Evolution, and Systematics, 2007. **38**(1): p. 515-540.

87. Elena, S.F. and R. Sanjuan, *Adaptive value of high mutation rates of RNA viruses: separating causes from consequences.* J Virol, 2005. **79**(18): p. 11555-8.

88. Paterson, S., et al., *Antagonistic coevolution accelerates molecular evolution.* Nature, 2010. **464**(7286): p. 275-8.

89. Sanjuan, R., et al., *Viral mutation rates.* J Virol, 2010. **84**(19): p. 9733-48.

90. Dimitrov, D.S., *Virus entry: Molecular mechanisms and biomedical applications.* Nature Reviews Microbiology, 2004. **2**(2): p. 109-122.

91. Grove, J. and M. Marsh, *The cell biology of receptor-mediated virus entry.* Journal of Cell Biology, 2011. **195**(7): p. 1071-1082.

92. Muckenthaler, M.U., B. Galy, and M.W. Hentze, *Systemic iron homeostasis and the iron-responsive element/iron-regulatory protein (IRE/IRP) regulatory network.* Annu Rev Nutr, 2008. **28**: p. 197-213.

93. Sallusto, F. and M. Baggiolini, *Chemokines and leukocyte traffic.* Nature Immunology, 2008. **9**(9): p. 949-952.

94. Miyoshi, J. and Y. Takai, *Nectin and nectin-like molecules: biology and pathology.* Am J Nephrol, 2007. **27**(6): p. 590-604.

95. Rajagopalan, L. and K. Rajarathnam, *Structural basis of chemokine receptor function - A model for binding affinity and ligand selectivity.* Bioscience Reports, 2006. **26**(5): p. 325-339.

96. Alcami, A., *Viral mimicry of cytokines, chemokines and their receptors.* Nature Reviews Immunology, 2003. **3**(1): p. 36-50.

97. Locati, M. and P.M. Murphy, *Chemokines and chemokine receptors: Biology and clinical relevance in inflammation and AIDS.* Annual Review of Medicine, 1999. **50**: p. 425-440.

98. Tripp, R.A., et al., *CX3C chemokine mimicry by respiratory syncytial virus G glycoprotein.* Nature Immunology, 2001. **2**(8): p. 732-738.

99. Jahan, S., et al., *HCV entry receptors as potential targets for siRNA-based inhibition of HCV.* Genetic vaccines and therapy, 2011. **9**: p. 15-15.

100. Alhoot, M.A., S.M. Wang, and S.D. Sekaran, *Inhibition of Dengue Virus Entry and Multiplication into Monocytes Using RNA Interference.* Plos Neglected Tropical Diseases, 2011. **5**(11).

101. Zhang, Y.M., et al., *Identification of the receptor binding domain of the mouse mammary tumor virus envelope protein.* Journal of Virology, 2003. **77**(19): p. 10468-10478.

102. Demogines, A., et al., *Dual Host-Virus Arms Races Shape an Essential Housekeeping Protein.* Plos Biology, 2013. **11**(5).

103. Kerr, S.A., et al., *Computational and Functional Analysis of the Virus-Receptor Interface Reveals Host Range Trade-Offs in New World Arenaviruses.* Journal of Virology, 2015. **89**(22): p. 11643-11653.

104. Kaelber, J.T., et al., *Evolutionary reconstructions of the transferrin receptor of Caniforms supports canine parvovirus being a re-emerged and not a novel pathogen in dogs.* PLoS Pathog, 2012. **8**(5): p. e1002666.

105. Zhang, J.C., et al., *Down-regulation of CXCR4 expression by SDF-KDEL in CD34(+) hematopoietic stem cells: An anti-human immunodeficiency virus strategy.* J Virol Methods, 2009. **161**(1): p. 30-7.

106. Wunder, T., et al., *Expression of the coxsackie adenovirus receptor in neuroendocrine lung cancers and its implications for oncolytic adenoviral infection.* Cancer Gene Therapy, 2013. **20**(1): p. 25-32.

107. Li, Y.M., et al., *Loss of adenoviral receptor expression in human bladder cancer cells: A potential impact on the efficacy of gene therapy.* Cancer Research, 1999. **59**(2): p. 325-330.

108. Carroll, S.B., *Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.* Cell, 2008. **134**(1): p. 25-36.

109. Wray, G.A., *The evolutionary significance of cis-regulatory mutations.* Nat Rev Genet, 2007. **8**(3): p. 206-16.

110. Siegal, M.L. and A. Bergman, *Waddington's canalization revisited: Developmental stability and evolution.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(16): p. 10528-10532.

111. Leclerc, R.D., *Survival of the sparsest: robust gene networks are parsimonious.* Molecular Systems Biology, 2008. **4**.

112. Begon, M., et al., *A clarification of transmission terms in host-microparasite models: numbers, densities and areas.* Epidemiology and Infection, 2002. **129**(1): p. 147-153.

113. Saenz, R.A. and H.W. Hethcote, *Competing species models with an infectious disease.* Mathematical Biosciences and Engineering, 2006. **3**(1): p. 219-235.

114. Keeling, M.J. and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. 2008: Princeton Univ. Press.

115. Holmes, E.C., *Evolutionary history and phylogeography of human viruses.* Annu Rev Microbiol, 2008. **62**: p. 307-28.

116. Magurran, A.E., *Measuring Biological Diversity*. 2004, Oxford, United Kingdom: Wiley-Blackwell.

117. Samson, M., et al., *Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene.* Nature, 1996. **382**(6593): p. 722-725.

118. Dean, M., et al., *Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene.* Science, 1996. **273**(5283): p. 1856-1862.

119. Elleder, D., et al., *Two different molecular defects in the Tva receptor gene explain the resistance of two tva(r) lines of chickens to infection by subgroup A avian sarcoma and leukosis viruses.* Journal of Virology, 2004. **78**(24): p. 13489-13500.

120. Cao, X., *Self-regulation and cross-regulation of pattern-recognition receptor signalling in health and disease.* Nat Rev Immunol, 2016. **16**(1): p. 35-50.

121. Takeuchi, O. and S. Akira, *Pattern recognition receptors and inflammation.* Cell, 2010. **140**(6): p. 805-20.

122. Akira, S. and K. Takeda, *Toll-like receptor signalling.* Nat Rev Immunol, 2004. **4**(7): p. 499-511.

123. O'Neill, L.A., *When signaling pathways collide: positive and negative regulation of toll-like receptor signal transduction.* Immunity, 2008. **29**(1): p. 12-20.

124. Rosenberger, C.M. and B.B. Finlay, *Phagocyte sabotage: disruption of macrophage signalling by bacterial pathogens.* Nat Rev Mol Cell Biol, 2003. **4**(5): p. 385-96.

125. Bowie, A.G. and L. Unterholzner, *Viral evasion and subversion of pattern-recognition receptor signalling.* Nat Rev Immunol, 2008. **8**(12): p. 911-22.

126. Honda, K. and T. Taniguchi, *IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors.* Nat Rev Immunol, 2006. **6**(9): p. 644-58.

127. Revill, P. and Z. Yuan, *New insights into how HBV manipulates the innate immune response to establish acute and persistent infection.* Antivir Ther, 2013. **18**(1): p. 1-15.

128. Correia, S., S. Ventura, and R.M. Parkhouse, *Identification and utility of innate immune system evasion mechanisms of ASFV.* Virus Res, 2013. **173**(1): p. 87-100.

129. Powell, P.P., L.K. Dixon, and R.M. Parkhouse, *An IkappaB homolog encoded by African swine fever virus provides a novel mechanism for downregulation of proinflammatory cytokine responses in host macrophages.* J Virol, 1996. **70**(12): p. 8527-33.

130. Park, I.K., et al., *The Axl/Gas6 pathway is required for optimal cytokine signaling during human natural killer cell development.* Blood, 2009. **113**(11): p. 2470-7.

131. Drayman, N., et al., *Pathogens use structural mimicry of native host ligands as a mechanism for host receptor engagement.* Cell Host Microbe, 2013. **14**(1): p. 63-73.

132. Ashour, J., et al., *NS5 of dengue virus mediates STAT2 binding and degradation.* J Virol, 2009. **83**(11): p. 5408-18.

133. Elde, N.C., et al., *Protein kinase R reveals an evolutionary model for defeating viral mimicry.* Nature, 2009. **457**(7228): p. 485-9.

134. Barber, M.F. and N.C. Elde, *Evolutionary biology: Mimicry all the way down.* Nature, 2013. **501**(7465): p. 38-9.

135. Sackton, T.B., et al., *Dynamic evolution of the innate immune system in Drosophila.* Nat Genet, 2007. **39**(12): p. 1461-8.

136. Harpur, B.A. and A. Zayed, *Accelerated evolution of innate immunity proteins in social insects: adaptive evolution or relaxed constraint?* Mol Biol Evol, 2013. **30**(7): p. 1665-74.

137. Fares, M.A., *The origins of mutational robustness.* Trends Genet, 2015. **31**(7): p. 373-81.

138. Enard, D., et al., *Viruses are a dominant driver of protein adaptation in mammals.* Elife, 2016. **5**.

139. Calderwood, M.A., et al., *Epstein-Barr virus and virus human protein interaction maps.* Proc Natl Acad Sci U S A, 2007. **104**(18): p. 7606-11.

140. Guo, Y., et al., *Hepatitis B viral core protein disrupts human host gene expression by binding to promoter regions.* BMC Genomics, 2012. **13**: p. 563.

141. Lefort, S., et al., *Binding of Kaposi's sarcoma-associated herpesvirus K-bZIP to interferon-responsive factor 3 elements modulates antiviral gene expression.* J Virol, 2007. **81**(20): p. 10950-60.

142. Wang, H. and W.S. Ryu, *Hepatitis B virus polymerase blocks pattern recognition receptor signaling via interaction with DDX3: implications for immune evasion.* PLoS Pathog, 2010. **6**(7): p. e1000986.